|  |
| --- |
| **ePROJECT DATA SCIENCE**  **(REPORT ON STUDENT MENTAL HEALTH ANALYSIS)** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Developed by**  **Members:**   |  |  |  | | --- | --- | --- | | **No.** | **Student Name** | **Student ID** | | 1 | Dương Ngọc Thảo | Student1626510 | | 2 | Nguyễn Hưng Thiên Đức | Student1521204 | | 3 | Nguyễn Thị Thu Thủy | Student1648795 |   **Class No.: DT2503L-ON**  **Start Date: 10/11/2025**  **End Date: 08/12/2025**  **Name of the Coordinator: Đào Mạnh Thắng**  **Date of Submission: 08/12/2025** |

**MỤC LỤC**

[**CHƯƠNG 1: GIỚI THIỆU VÀ QUY TRÌNH THỰC HIỆN 4**](#_heading=h.ld46cp6w8yc2)

[1.1. Lý do chọn đề tài 4](#_heading=h.uceg6732sz2m)

[1.2. Mục tiêu nghiên cứu 4](#_heading=h.4y67ucrt9oji)

[1.3. Phạm vi nghiên cứu 5](#_heading=h.der0x9doaiiv)

[1.4. Phương pháp nghiên cứu 5](#_heading=h.2mtt56ayw1xi)

[1.4.1. Phương pháp Thu thập và Lưu trữ Dữ liệu 5](#_heading=h.d0plmg8b2u0p)

[1.4.2. Phương pháp Xử lý và Lập Mô hình (Python) 6](#_heading=h.ft6prmixy2ci)

[1.4.3. Phương pháp Trực quan hóa (Power BI) 6](#_heading=h.q0bfqsrjrleg)

[1.5. Ý nghĩa khoa học và thực tiễn 6](#_heading=h.yx0lig902ad3)

[**CHƯƠNG 2: TRÍCH XUẤT VÀ TIỀN XỬ LÝ DỮ LIỆU VỚI PYTHON 7**](#_heading=h.fo09dwvs75bt)

[2.1. Mục tiêu chương 7](#_heading=h.4qyqpr6iw669)

[2.2. Trích xuất Dữ liệu và Đánh giá Cấu trúc Ban đầu 7](#_heading=h.fmafu4ru4f8w)

[2.2.1. Nguồn dữ liệu và Quy trình Trích xuất (Extract) 7](#_heading=h.vritti8ut8p4)

[2.2.2. Tải và Đánh giá Cấu trúc Dữ liệu Thô (Load & Check) 7](#_heading=h.c0v9ql4o6lph)

[2.3. Kiểm tra và Xử lý Dữ liệu 10](#_heading=h.k91ycowj4009)

[2.3.1. Kiểm tra dữ liệu 10](#_heading=h.iyq0efk43qye)

[2.3.2. Xử lý dữ liệu 14](#_heading=h.axmg1h4wm2h1)

[2.3.3. Kiểm tra lại dữ liệu sau khi làm sạch 15](#_heading=h.19ufhsleis0g)

[2.4. Chuẩn hóa và Mã hóa Dữ liệu (Normalization & Encoding) 17](#_heading=h.dlcgqjhppb8u)

[2.4.1. Chuẩn hóa Dữ liệu Định tính (Categorical Values) 17](#_heading=h.qx378imsdxmq)

[2.4.2. Kiểm tra lại dữ liệu sau khi làm sạch 18](#_heading=h.utk8s8fqeo49)

[2.5. Hoàn tất Tiền xử lý và Xuất Dữ liệu Sạch 20](#_heading=h.jutb9n93j648)

[**CHƯƠNG 3: LƯU TRỮ VÀ PHÂN TÍCH DỮ LIỆU 23**](#_heading=h.2zjv3nsxry3w)

[3.1. Lưu trữ dữ liệu trên MongoDB 23](#_heading=h.v69yt9t7dant)

[3.2. Phương pháp phân tích 23](#_heading=h.9skr3y4r8jmc)

[3.3 Phân tích dữ liệu chuyên sâu 23](#_heading=h.97gtidymw15l)

[3.3.1. Trình độ học vấn và mức độ stress, lo âu trước kỳ thi 24](#_heading=h.ajl84wpqxsda)

[3.3.2. Phân tích mức độ stress và thời gian ngủ 27](#_heading=h.leeukk3e5jc9)

[3.3.4. Nhóm chưa cải thiện học tập (Academic Performance Change = "Declined") 33](#_heading=h.opqtt3cit1mw)

[3.3.5. Nhóm cải thiện học tập (Academic Performance Change = "Improved") 35](#_heading=h.2u6q6gje2vbg)

[3.3.6. Nhóm nguy cơ cực cao 37](#_heading=h.4fvvmi40vfd1)

[3.4. Tối ưu hiệu năng bằng Index trong MongoDB 39](#_heading=h.wd1m01gwo9nx)

[3.5. Kết luận chương 3 40](#_heading=h.e8y2o1qcw2ui)

[**CHƯƠNG 4 : TRỰC QUAN HÓA DỮ LIỆU 41**](#_heading=h.oo3628njevv7)

[4.1. Mục tiêu của phần phân tích trên Power BI 41](#_heading=h.dvqn4bm1c2uy)

[4.2. Mô tả cấu trúc và thiết kế Dashboard 42](#_heading=h.kbre9nm1r31c)

[**4.3. Insight chính (Key Findings) 45**](#_heading=h.th2hleg62e1)

[4.3.1. Nhóm sinh viên lớn tuổi (~35 tuổi) là ngoại lệ với thời gian ngủ cao và thời gian màn hình thấp 45](#_heading=h.cxk2apxblb99)

[4.3.2. Giới tính “other” độ tuổi ảnh hưởng đến stress 46](#_heading=h.qyl2h6yv14bj)

[4.3.3. Cả thời gian sử dụng màn hình và hoạt động thể chất đều tăng cùng với mức độ Stress. 46](#_heading=h.f8rbxgatti1g)

[4.3.4. Hoạt động thể chất càng cao thì Stress Level càng tăng — cho thấy sinh viên stress nặng tập thể dục nhiều hơn (mang tính đối phó) 47](#_heading=h.q29d13ehbh7n)

[4.3.5. Khả năng vượt qua các tác động ảnh hưởng đến tinh thần – Phân tích Level 2 47](#_heading=h.64ccfrfkcid9)

[4.3.6. Mô hình và con số báo động ở phân cụm nhóm At-Risk (33,53%) cần hỗ trợ thực thi 48](#_heading=h.dl6eg5f583md)

[4.4. Kể chuyện dữ liệu (Data Storytelling) 50](#_heading=h.9el3g8j19v92)

[4.5. Kết luận & Khuyến nghị 51](#_heading=h.z7w6np6ar754)

# CHƯƠNG 1: GIỚI THIỆU VÀ QUY TRÌNH THỰC HIỆN

## 1.1. **Lý do chọn đề tài**

Trong bối cảnh xã hội hiện nay, sức khỏe tinh thần ngày càng được xem là yếu tố quan trọng quyết định đến sự phát triển toàn diện của con người, đặc biệt là đối tượng sinh viên – những người đang trong giai đoạn học tập, định hướng nghề nghiệp và hình thành nhân cách. Áp lực học tập, kỳ vọng của gia đình, khó khăn tài chính, cũng như ảnh hưởng của mạng xã hội đã khiến nhiều sinh viên rơi vào trạng thái **căng thẳng (stress)**, **lo âu (anxiety)** hoặc **trầm cảm (depression)**.

Với sự phát triển mạnh mẽ của phân tích dữ liệu (data analytics), việc thu thập và phân tích dữ liệu sức khỏe tinh thần của sinh viên sẽ giúp nhận diện sớm các nhóm có nguy cơ cao, từ đó đưa ra các biện pháp hỗ trợ kịp thời. Chính vì vậy, đề tài “Phân tích tình hình sức khỏe tinh thần sinh viên” được lựa chọn với mong muốn ứng dụng các công cụ hiện đại như Python, MongoDB và Power BI để mô hình hóa, phân tích và trực quan hóa dữ liệu, góp phần nâng cao nhận thức và chất lượng hỗ trợ sinh viên.

## **1.2. Mục tiêu nghiên cứu**

**Mục tiêu chung**

Phân tích chuyên sâu dữ liệu sức khỏe tinh thần của sinh viên để nhận diện các yếu tố nguy cơ chính (thời gian sử dụng màn hình, hoạt động thể chất, giấc ngủ...) ảnh hưởng đến trạng thái tâm lý, từ đó cung cấp cơ sở dữ liệu định lượng (data-driven insights) hỗ trợ công tác đánh giá và xây dựng các chương trình can thiệp hiệu quả tại trường học.

**Mục tiêu cụ thể**

Quản lý dữ liệu: Xây dựng và triển khai hệ thống lưu trữ dữ liệu phi cấu trúc bằng MongoDB và đảm bảo chất lượng dữ liệu ban đầu.

Tiền xử lý và EDA: Sử dụng Python để làm sạch, chuẩn hóa dữ liệu, sau đó thực hiện Exploratory Data Analysis (EDA) để khám phá mối tương quan.

Xây dựng Mô hình: Phân cụm sinh viên theo mức độ stress (K-means) và xây dựng mô hình dự đoán xác suất sinh viên có nguy cơ stress cao (Hồi quy Logistic/Cây Quyết định).

Trực quan hóa: Xây dựng Power BI Dashboard tương tác để trực quan hóa kết quả phân tích và các insight quan trọng.

Đề xuất giải pháp: Đưa ra các đề xuất hành động dựa trên kết quả mô hình nhằm cải thiện sức khỏe tinh thần sinh viên.

## **1.3. Phạm vi nghiên cứu**

**Đối tượng nghiên cứu**

Sinh viên thuộc nhiều độ tuổi và cấp bậc đào tạo khác nhau đang tham gia các hình thức học tập trực tuyến.

### Dữ liệu nghiên cứu

Nghiên cứu sử dụng tập dữ liệu liên quan đến hành vi và sức khỏe tinh thần của sinh viên, bao gồm:

* Thời gian sử dụng thiết bị điện tử
* Thời gian ngủ
* Mức độ hoạt động thể chất
* Mức độ stress
* Mức độ lo âu trước kỳ thi
* Giới tính, độ tuổi và một số đặc điểm nhân khẩu học khác

### Phạm vi công nghệ sử dụng

* **Python**: Dùng để tiền xử lý dữ liệu, phân tích thống kê và xây dựng mô hình dự đoán (sử dụng các thư viện Pandas, Scikit-learn, Scipy).
* **MongoDB**: Lưu trữ dữ liệu thô và hỗ trợ kiểm tra, đánh giá chất lượng dữ liệu ban đầu.
* **Power BI**: Trực quan hóa dữ liệu, xây dựng dashboard và trình bày kết quả phân tích một cách trực quan.

## **1.4. Phương pháp nghiên cứu**

Đề tài sử dụng kết hợp các phương pháp, tập trung vào quy trình phân tích dữ liệu chuyên nghiệp:

### 1.4.1. Phương pháp Thu thập và Lưu trữ Dữ liệu

Dữ liệu được trích xuất từ hệ thống Kangle.

Sử dụng MongoDB để lưu trữ dữ liệu theo dạng NoSQL.

### 1.4.2. Phương pháp Xử lý và Lập Mô hình (Python)

Làm sạch dữ liệu: Xử lý null, duplicate, dtype, outliers.

Phân tích mô tả: Thống kê mô tả, EDA, Phân tích tương quan.

Phân tích nâng cao: Phân cụm (K-means) và Phân loại (Logistic Regression, Decision Tree).

### 1.4.3. Phương pháp Trực quan hóa (Power BI)

Xây dựng Biểu đồ phân bố, so sánh giữa các nhóm sinh viên.

Thiết lập Dashboard tổng hợp và trình bày Insight chuyên sâu.

## **1.5. Ý nghĩa khoa học và thực tiễn**

**a. Ý nghĩa khoa học**

Minh chứng cho việc kết hợp hệ thống NoSQL (MongoDB) + Python + BI (Power BI) trong phân tích dữ liệu giáo dục.

Góp phần xây dựng hệ thống phân tích dữ liệu tự động từ lưu trữ → xử lý → báo cáo.

**b. Ý nghĩa thực tiễn**

Giúp nhà trường nhận diện nhóm sinh viên có nguy cơ stress cao sẽ ảnh hưởng đến kết quả học tập thông qua mô hình dự đoán.

Đưa ra gợi ý cải thiện môi trường học tập và sức khỏe tinh thần dựa trên các yếu tố nguy cơ được tìm thấy.

Tạo cơ sở dữ liệu vững chắc cho các chương trình chăm sóc tâm lý học đường.

# CHƯƠNG 2: TRÍCH XUẤT VÀ TIỀN XỬ LÝ DỮ LIỆU VỚI PYTHON

## **2.1. Mục tiêu chương**

Mục tiêu của chương này là thực hiện quy trình **Trích xuất (Extract)** dữ liệu thô từ nền tảng Kangle, và **Tiền xử lý, chuẩn hóa (Transform)** dữ liệu bằng các công cụ Python. Việc này nhằm tạo ra bộ dữ liệu **sạch, có chất lượng cao**, loại bỏ dữ liệu trùng lặp, thiếu hoặc bất thường, đồng thời chuẩn hóa các biến định tính, sẵn sàng cho **phân tích và xây dựng mô hình ở Chương 3**.

* **Kết quả đạt được:**

Hoàn thành quy trình tiền xử lý dữ liệu cơ bản.

Tạo ra bộ dữ liệu sạch với 1005 bản ghi, chuẩn hóa các biến định tính, tối ưu cho các bước phân tích nâng cao.

## **2.2. Trích xuất Dữ liệu và Đánh giá Cấu trúc Ban đầu**

### 2.2.1. Nguồn dữ liệu và Quy trình Trích xuất (Extract)

**Nguồn dữ liệu:** Nền tảng quản lý học tập trực tuyến **Kangle** cung cấp dữ liệu về: <https://www.kaggle.com/datasets/utkarshsharma11r/student-mental-health-analysis>

**Thông tin sinh viên:** Name, Gender, Age, Education Level

**Hoạt động học tập:** Screen Time (giờ/ngày), số giờ online

**Trạng thái tâm lý:** Stress Level, Anxious Before Exams, Academic Performance Change (có thể kết hợp khảo sát bổ sung)

* **Quy trình trích xuất:**

Truy cập Kangle với tài khoản cá nhân.

Xuất dữ liệu sinh viên theo phân tích *Student Mental Health Analysis During Online Learning*.

Lưu dữ liệu dưới định dạng CSV để tải vào Python.

* **Kết quả đạt được:**

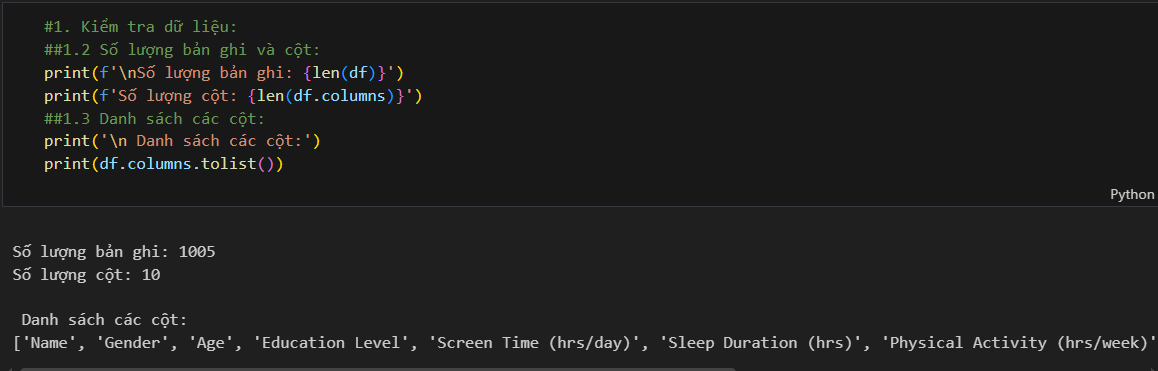
Trích xuất thành công N bản ghi dữ liệu thô với đầy đủ trường thông tin, sẵn sàng để phân tích trong môi trường Python.

### 2.2.2. Tải và Đánh giá Cấu trúc Dữ liệu Thô (Load & Check)

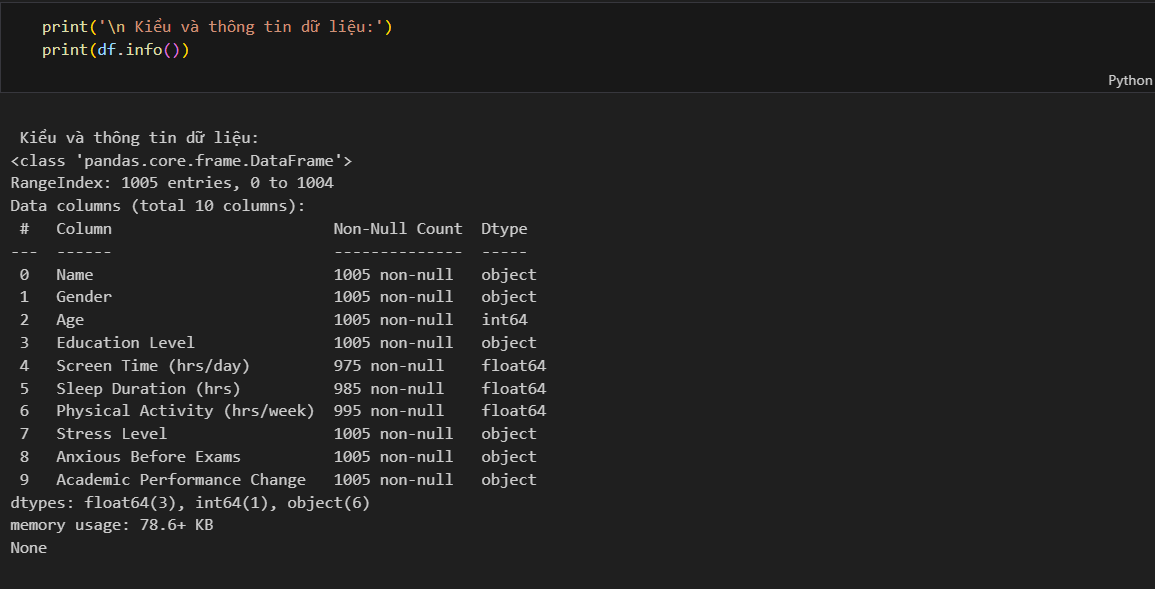
**Tải dữ liệu:** Sử dụng thư viện **pandas** để đọc tệp CSV vào **DataFrame**.

**Đánh giá cấu trúc ban đầu:**

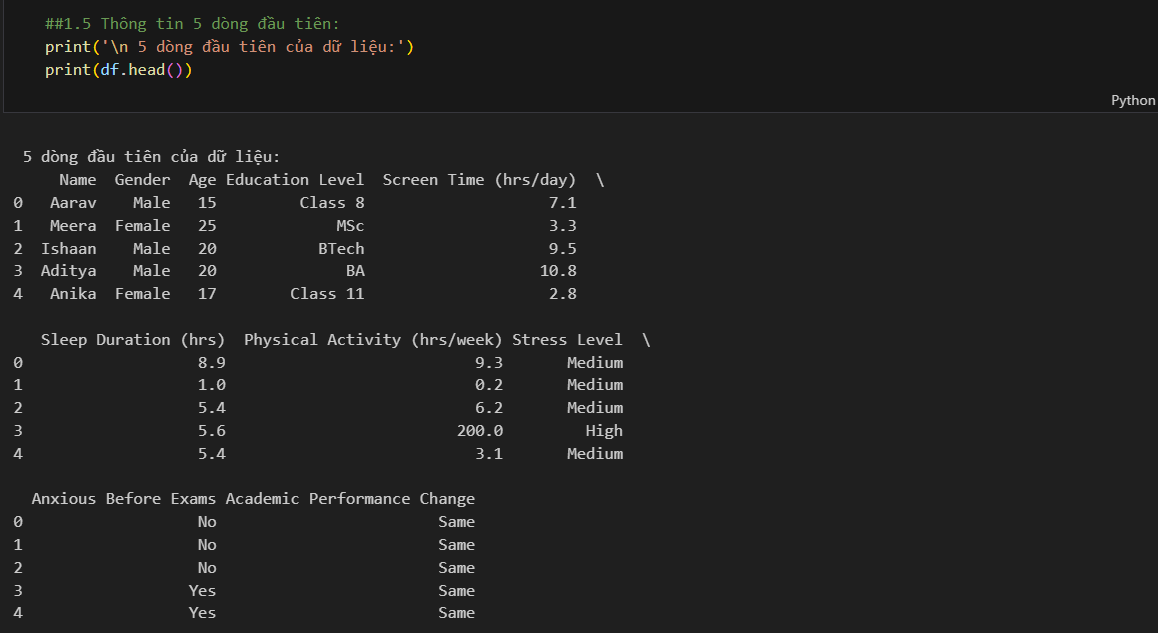
* Xem số lượng bản ghi và danh sách các cột.



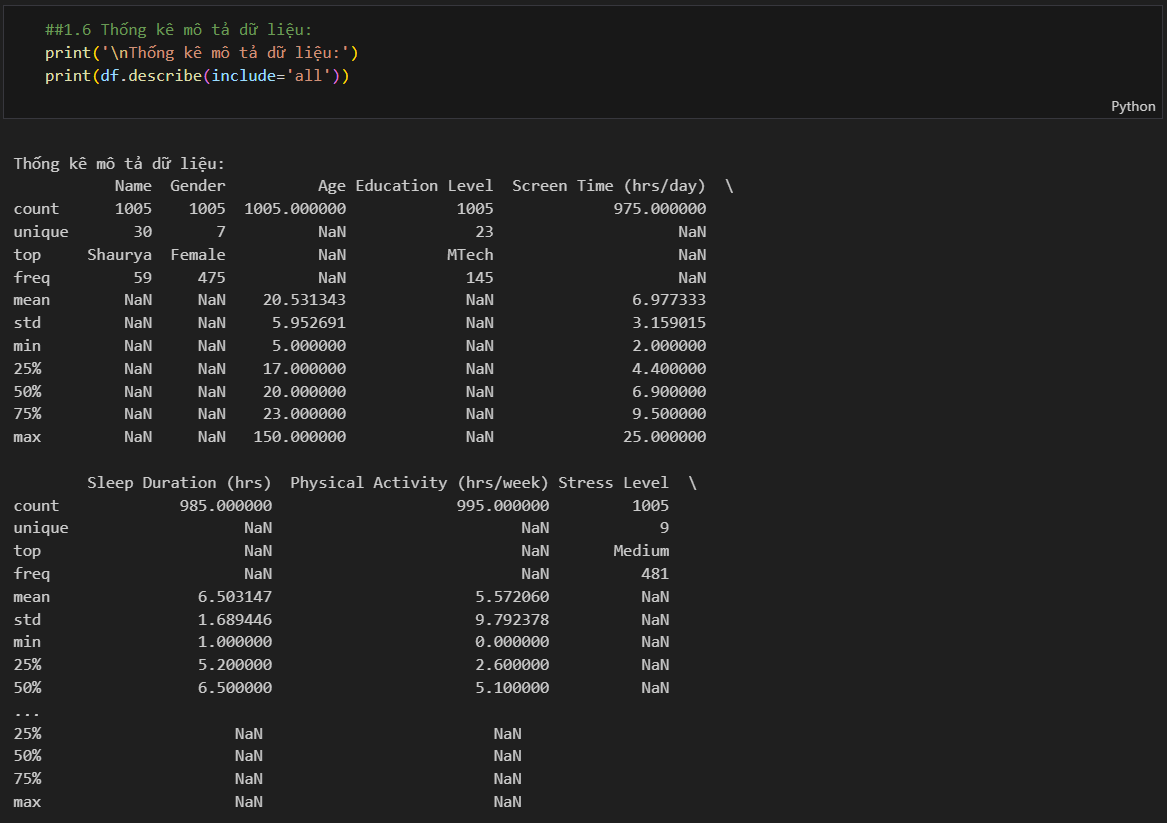
* Xem thông tin chi tiết về kiểu dữ liệu và số lượng giá trị không null.



* Hiển thị 5 dòng đầu tiên của dữ liệu.



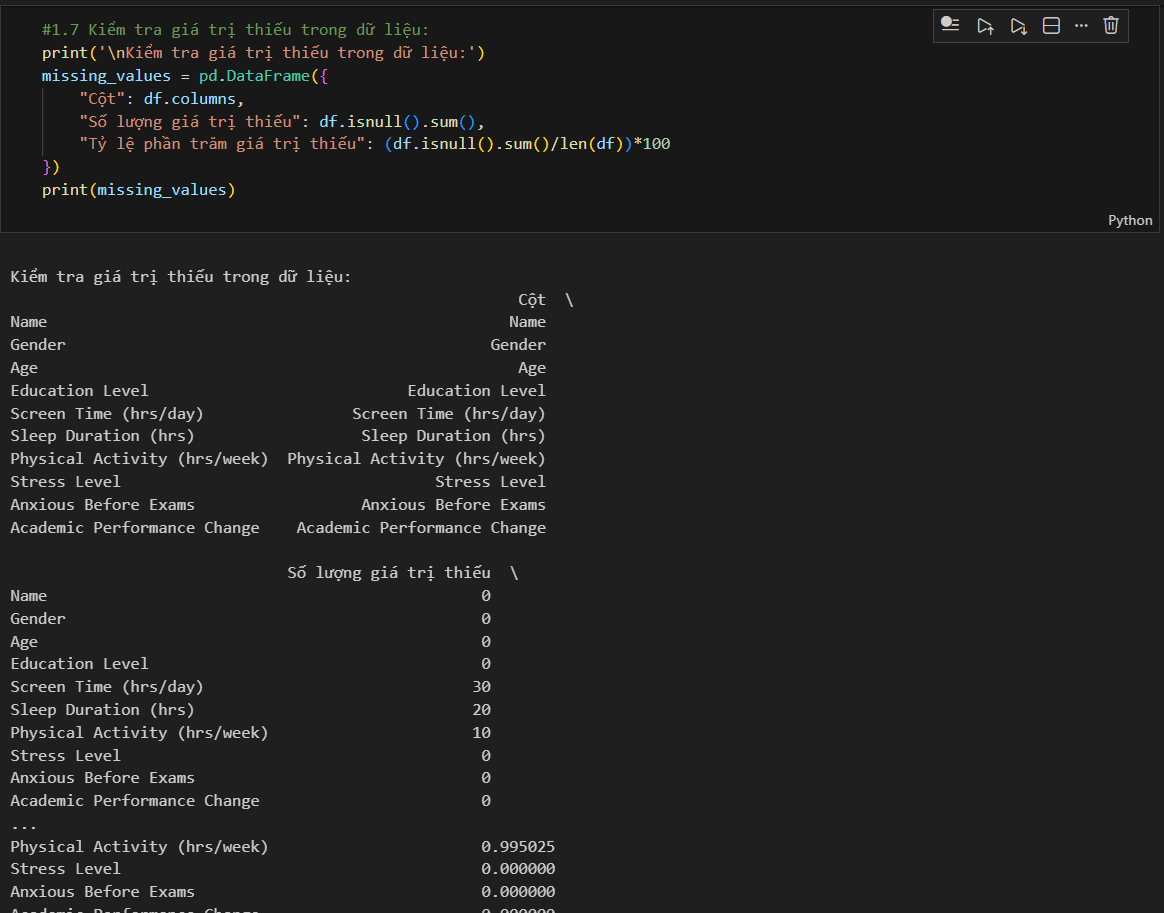
* Thống kê mô tả các biến định lượng để nắm bắt phân phối và giá trị trung tâm.



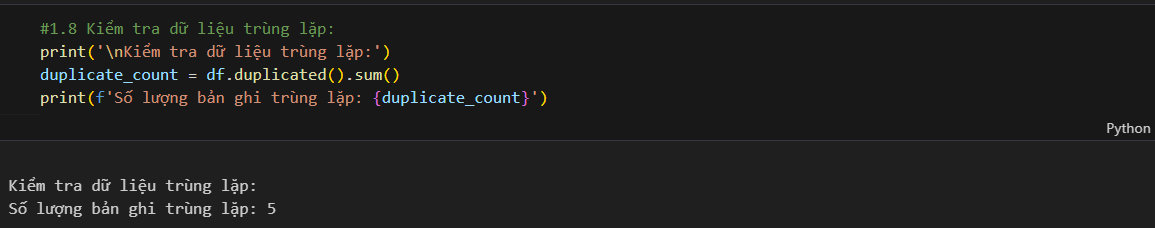
## **2.3. Kiểm tra và Xử lý Dữ liệu**

### 2.3.1. Kiểm tra dữ liệu

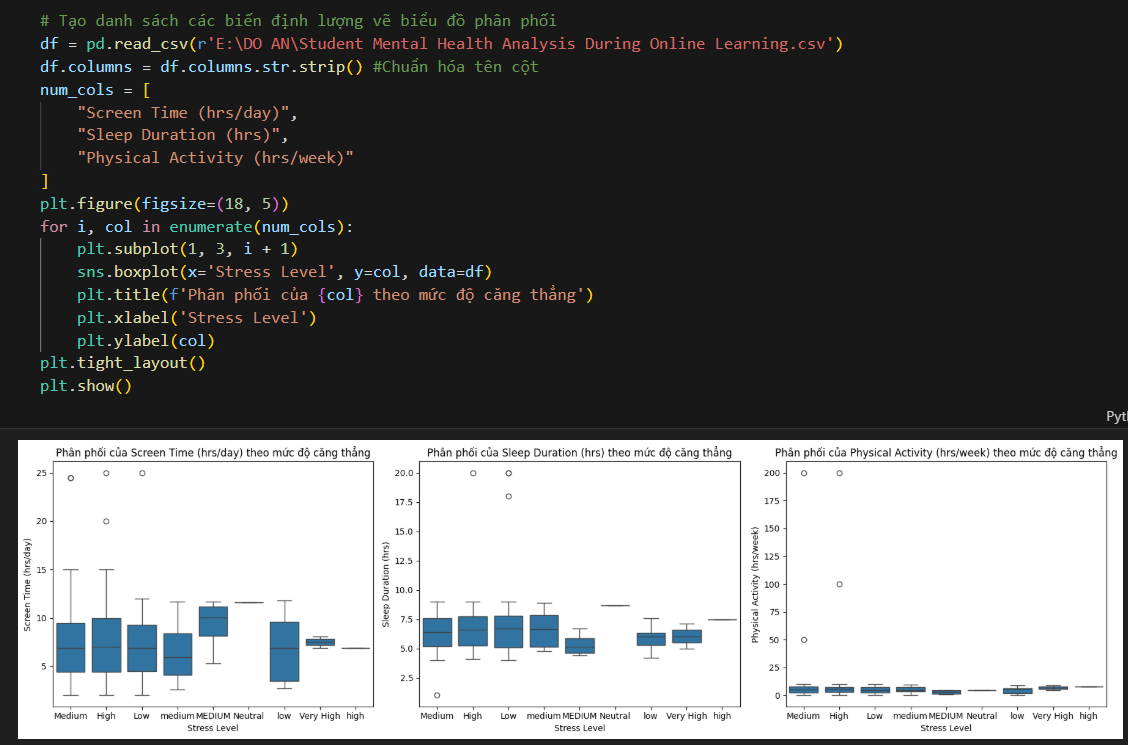
**- Dữ liệu thiếu:** Kiểm tra sự tồn tại và tỷ lệ giá trị missing(null) trong các cột



* **Dữ liệu trùng lặp:** Xác định các bản ghi trùng lặp hoàn toàn



* **Tạo danh sách các biến định lượng vẽ biểu đồ phân phối**



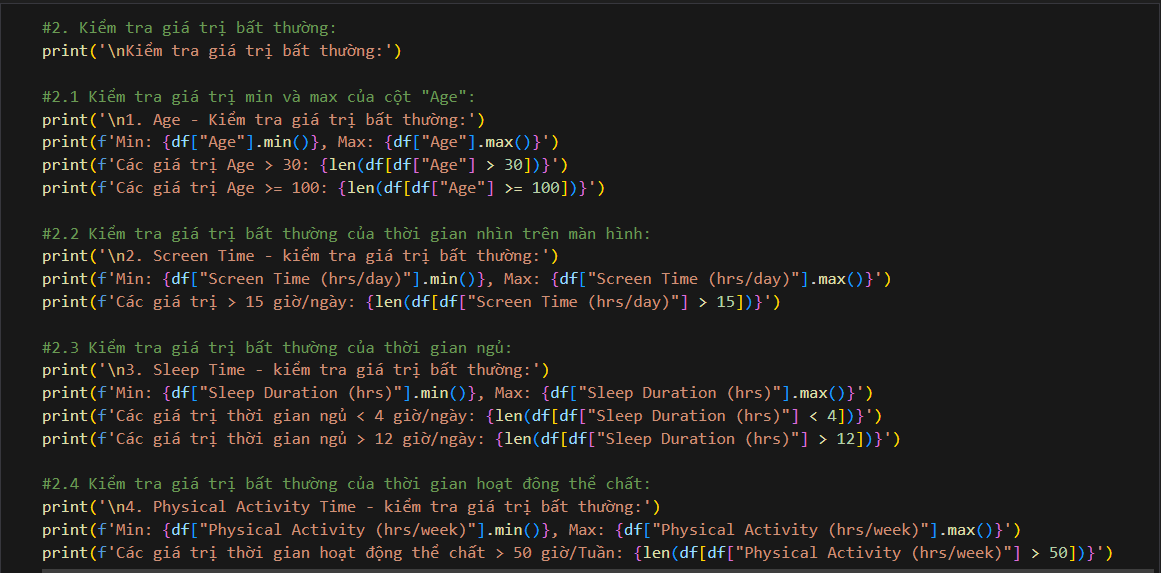
* **Kiểm tra dữ liệu bất thường (Ngưỡng Logic):** Xác định các giá trị nằm ngoài phạm vi logic

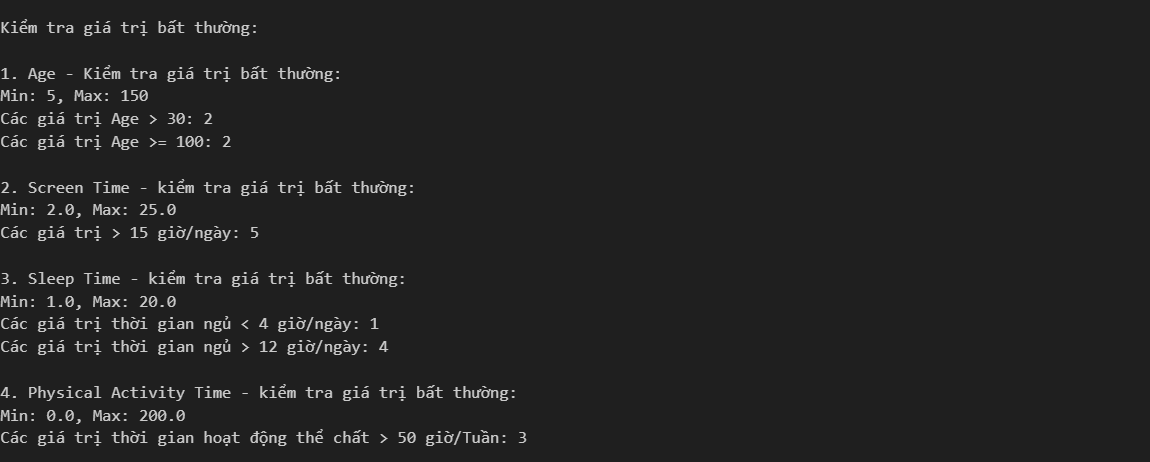
Age > 100 (lỗi nhập liệu).

Screen Time > 15 giờ/ngày.

Thời gian ngủ < 4 giờ/ngày.

Thời gian hoạt động thể chất > 50 giờ/tuần.



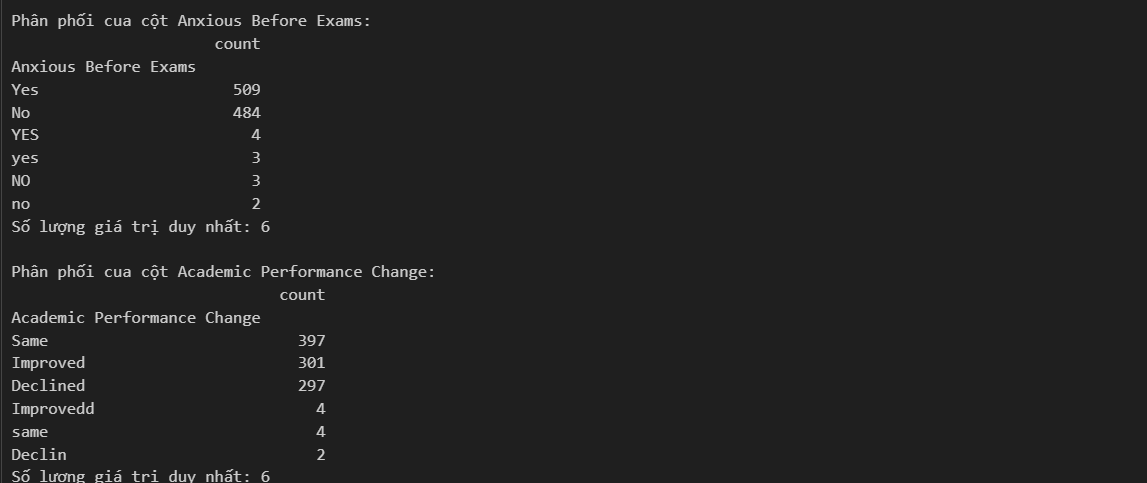


* **Kiểm tra Phân phối các biến định lượng**

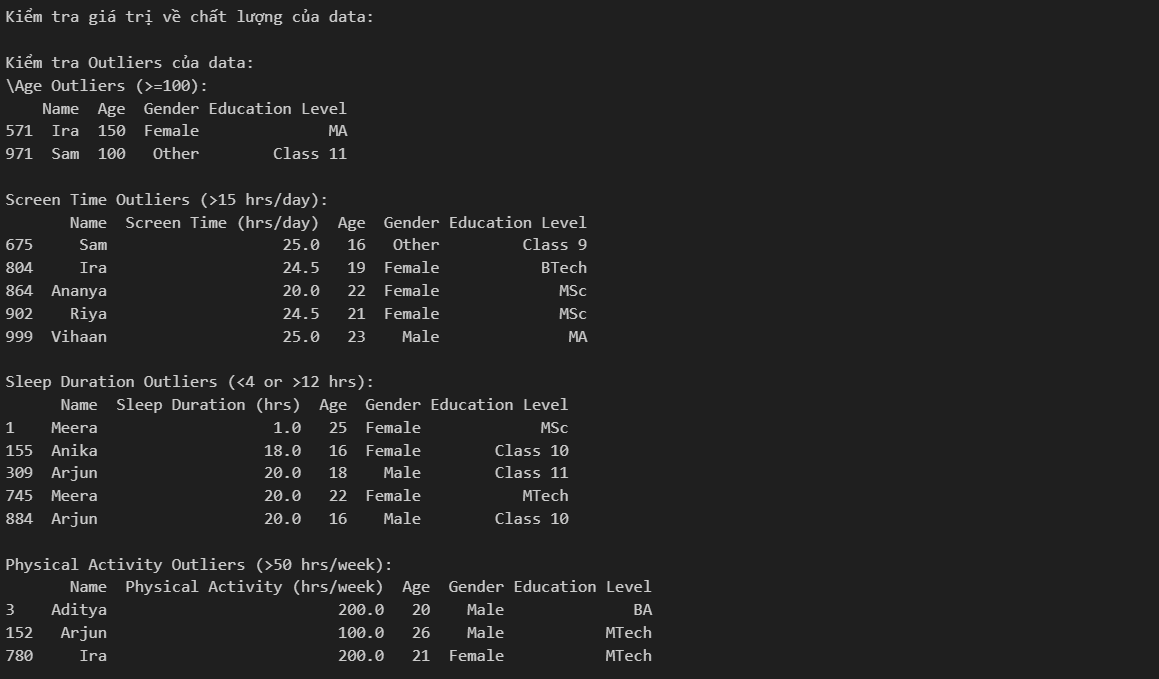




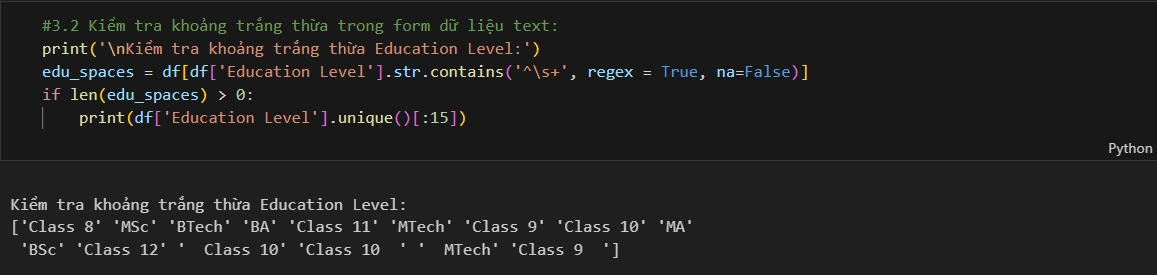




* **Kiểm tra chất lượng data:**



* **Kiểm tra khoảng trắng thừa trong form text:**



### 2.3.2. Xử lý dữ liệu

**- Xử lý dữ liệu thiếu:**



* **Xử lý giá trị bất thường:**

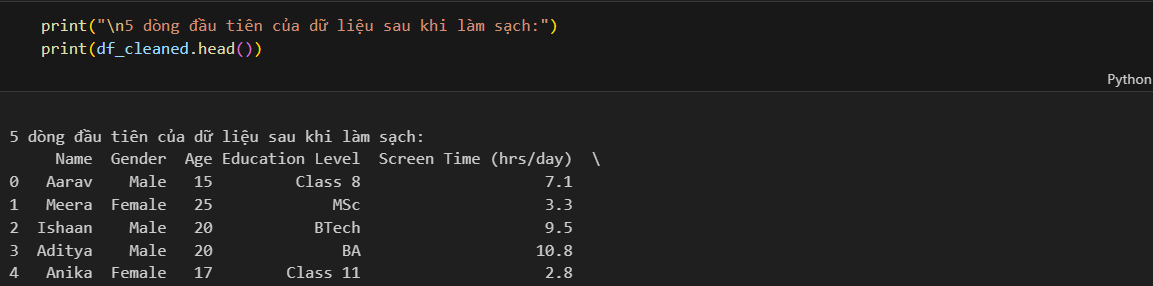


### 2.3.3. Kiểm tra lại dữ liệu sau khi làm sạch

* **Kiểm tra tổng quan dữ liệu làm sạch:**

****

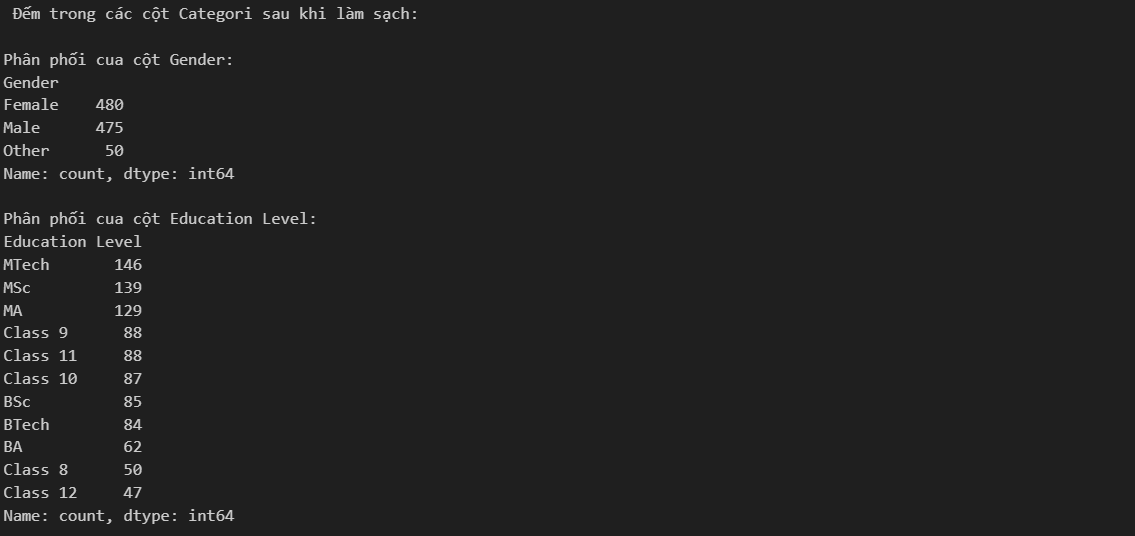
* **Kiểm tra 5 dòng đầu tiên:**

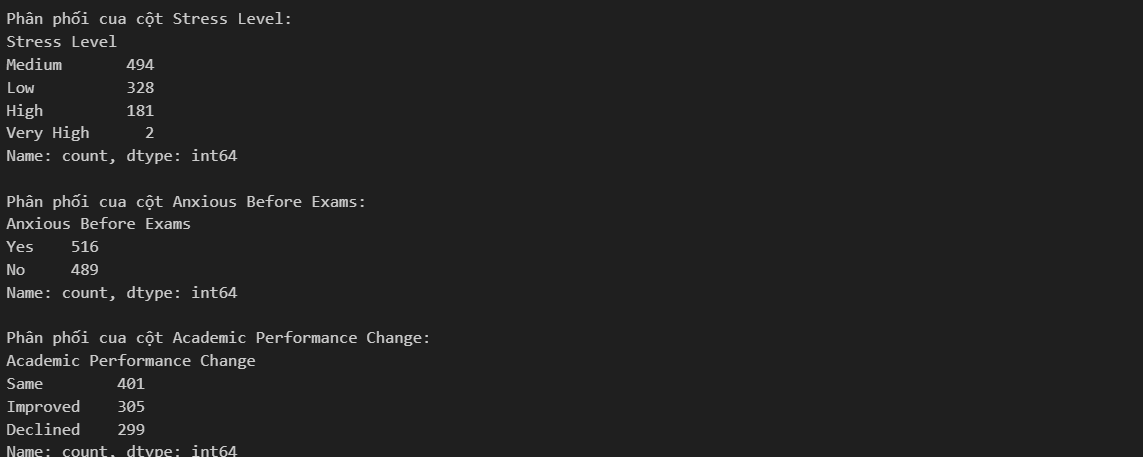
****

* **Thống kê mô tả dữ liệu sau khi làm sạch:**

****

* **Đếm các cột Category sau khi làm sạch:**





## **2.4. Chuẩn hóa và Mã hóa Dữ liệu (Normalization & Encoding)**

### 2.4.1. Chuẩn hóa Dữ liệu Định tính (Categorical Values)

**- Chuẩn hóa cột định tính:**

Gender

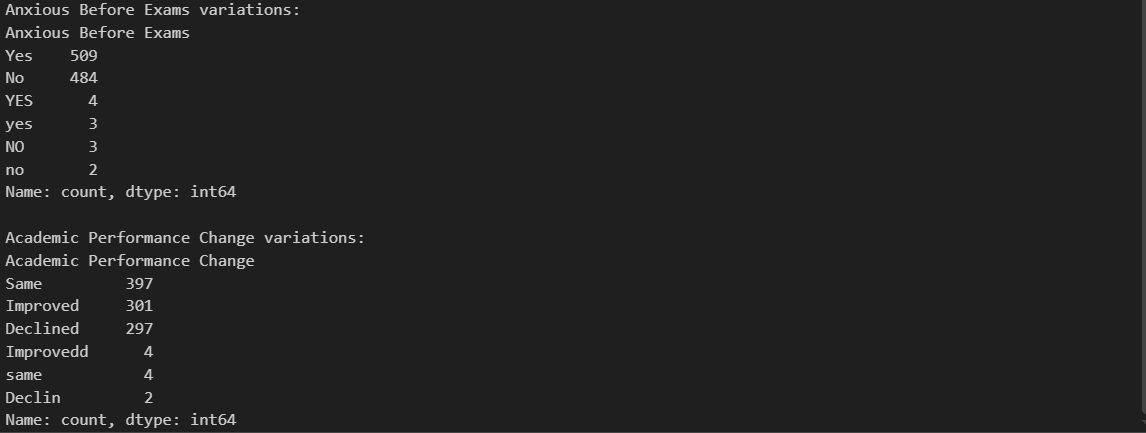
Stress Level

Anxious Before Exams

Academic Performance Change

Education Level





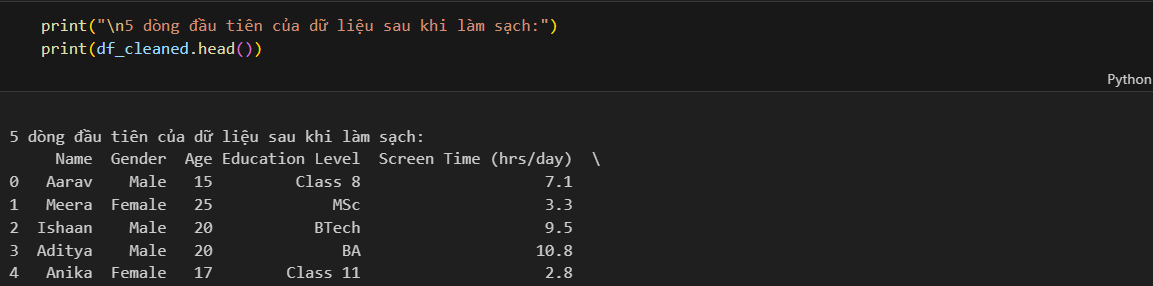
**Kết quả đạt được:** Các biến định tính đã được chuẩn hóa, sẵn sàng cho mã hóa số

### 2.4.2. Kiểm tra lại dữ liệu sau khi làm sạch

**- Kiểm tra tổng quan dữ liệu làm sạch:**



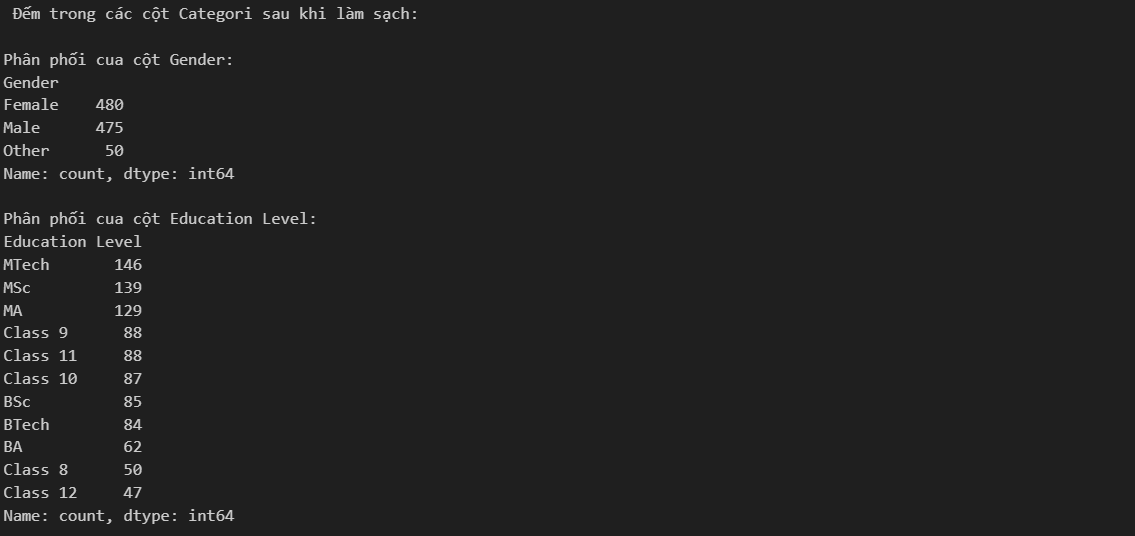
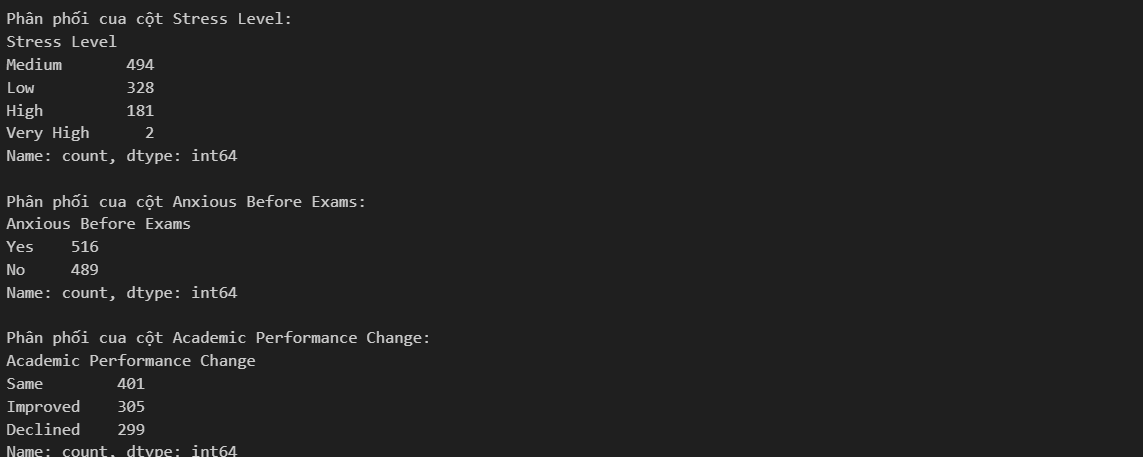
* **Kiểm tra 5 dòng đầu tiên:**



* **Thống kê mô tả dữ liệu sau khi làm sạch:**



* **Đếm các cột Category sau khi làm sạch:**

 .

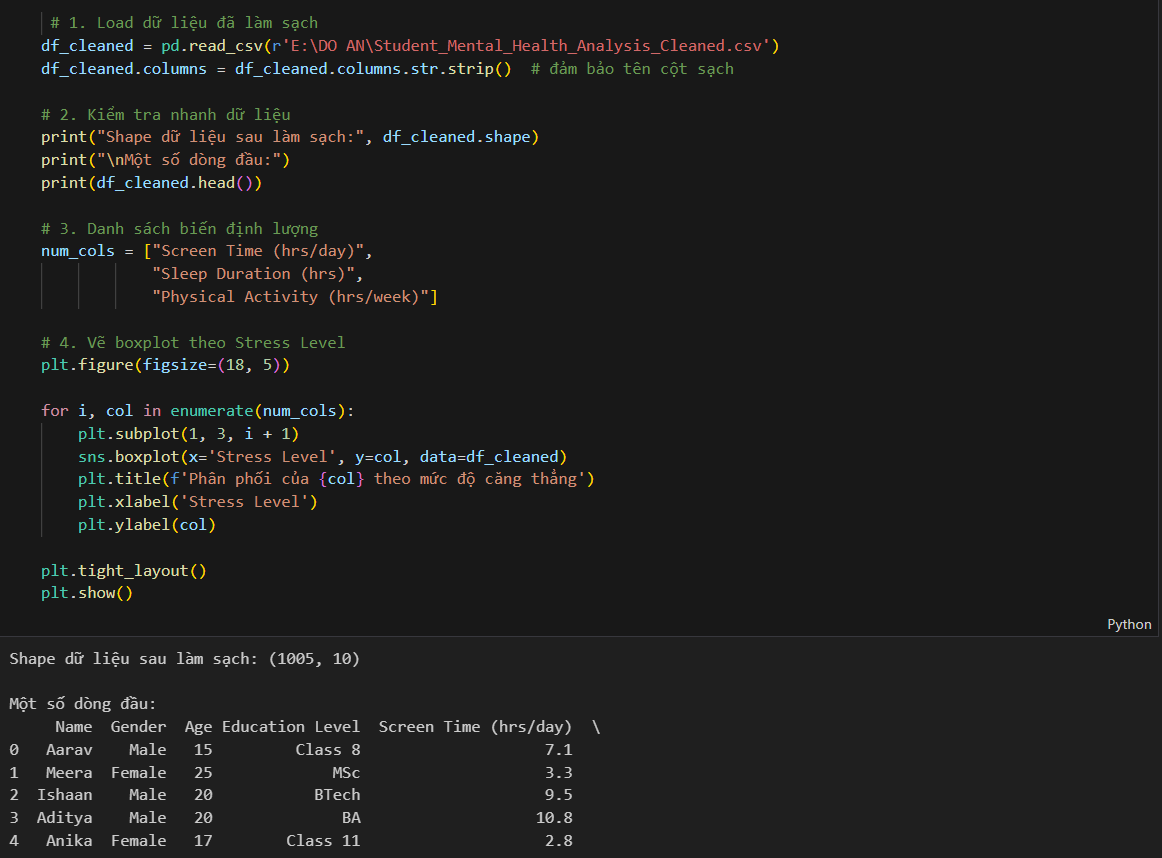
## **2.5. Hoàn tất Tiền xử lý và Xuất Dữ liệu Sạch**

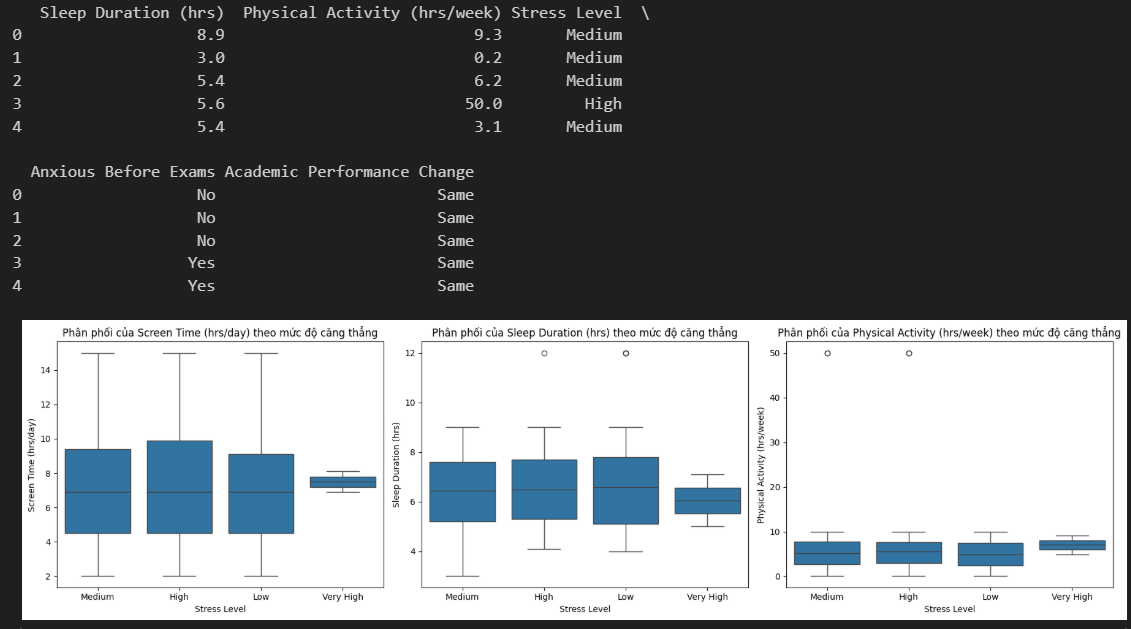
**- Kiểm tra cuối cùng:**

Đảm bảo bộ dữ liệu không còn missing, trùng lặp, outliers.

Kiểm tra tính đồng nhất và chuẩn hóa các biến.

Tạo danh sách các biến định lượng, vẽ biểu đồ phân phối sau khi làm sạch





**Xuất dữ liệu CSV:**

Lưu dữ liệu đã làm sạch và chuẩn hóa cuối cùng vào file **Cleaned\_Data.csv**.

Đây là nguồn dữ liệu chính thức cho các chương tiếp theo (Phân tích, Mô hình, và lưu trữ vào MongoDB).

**Thành tựu chương:**

Bộ dữ liệu sạch, chất lượng cao, đã chuẩn hóa và mã hóa thành công.

Sẵn sàng cho phân tích chuyên sâu và xây dựng mô hình ở Chương 3.

# CHƯƠNG 3: LƯU TRỮ VÀ PHÂN TÍCH DỮ LIỆU

## **3.1. Lưu trữ dữ liệu trên MongoDB**

Dữ liệu khảo sát sức khỏe tinh thần của sinh viên được lưu trữ trong hệ quản trị cơ sở dữ liệu NoSQL MongoDB. Bộ dữ liệu đã được làm sạch, chuẩn hóa và nhập vào collection như sau:

**Database:** DO\_AN\_CUOI\_KY

**Collection:** Student\_Mental\_Health\_Analysis\_Cleaned

MongoDB được lựa chọn vì:

Khả năng xử lý dữ liệu phi cấu trúc linh hoạt.

Hiệu quả trong việc phân tích dữ liệu mức độ lớn nhờ Aggregation Pipeline.

Hỗ trợ tốt cho các tác vụ thống kê dạng nhóm, lọc và tính toán trung bình.

## **3.2. Phương pháp phân tích**

Chương này sử dụng **Aggregation Pipeline của MongoDB** để phân tích dữ liệu theo 6 góc nhìn (insight) quan trọng nhằm đánh giá sức khỏe tinh thần và hành vi của sinh viên.

**Các phép toán chủ đạo trong Aggregation Pipeline:**

**$match:** lọc dữ liệu theo điều kiện cụ thể.

**$group:** tổng hợp dữ liệu theo nhóm.

**$project:** tạo hoặc tính toán giá trị mới.

**$bucket:** phân nhóm dữ liệu theo khoảng (ví dụ: độ tuổi).

**$avg, $sum:** tính toán giá trị trung bình, tổng.

**$sort:** sắp xếp kết quả theo thứ tự tăng/giảm.

Phương pháp này đảm bảo phân tích từ tổng quan đến chi tiết, khai thác được mối quan hệ giữa các biến định lượng và định tính, đồng thời tối ưu hóa hiệu năng truy vấn cho bộ dữ liệu lớn.

## **3.3 Phân tích dữ liệu chuyên sâu**

Tổng quan dữ liệu:

**Aggregation Pipeline:**

**$group:** Tổng hợp các chỉ số cần thiết khi sử dụng để so sánh các chỉ số như: Trung bình và trung vị của thời gian màn hình, thời gian ngủ, thời gian hoạt động.

**Sau khi chạy Pipeline sẽ có kết quả như sau:**

|  |
| --- |
| C:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\5A7EC61C.tmp |

### 3.3.1. Trình độ học vấn và mức độ stress, lo âu trước kỳ thi

**- Mục tiêu:**  
Xác định và đánh giá mối liên hệ giữa trình độ học vấn với tỷ lệ sinh viên thuộc các mức độ Stress khác nhau và lo âu trước kỳ thi.

**Aggregation Pipeline:**

**$group:** fore Exam.

**$project:** Tính tổng hợp số lượng sinh viên theo Education Level, Stress Level và Anxious BE tỷ lệ phần trăm sinh viên trong mỗi nhóm.

**$sort:** Sắp xếp kết quả theo Education Level.

**Sau khi chạy Pipeline sẽ có kết quả như sau:**

|  |  |
| --- | --- |
| C:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\750619E3.tmpC:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\811610E9.tmp | C:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\C7B64BDF.tmpC:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\FC7DA045.tmp |

* **Kết quả phân tích:**

**Unit: %**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Education Level** | **Number of** | **Anxious** | **Stress High** | **Stress Medium** | **Stress Low** |
| **BA** | 62 | 58,06 | 27,42 | 51,61 | 20,97 |
| **BSc** | 85 | 60,00 | 16,47 | 50,59 | 32,94 |
| **BTech** | 84 | 55,95 | 25,00 | 44,05 | 30,95 |
| **Class 12** | 47 | 48,93 | 17,02 | 42,55 | 40,43 |
| **Class 11** | 88 | 50,00 | 18,18 | 52,27 | 29,54 |
| **Class 10** | 87 | 49,42 | 16,09 | 49,42 | 34,48 |
| **Class 9** | 88 | 50,00 | 13,63 | 45,45 | 40,91 |
| **Class 8** | 50 | 46,00 | 18,00 | 54,00 | 28,00 |
| **MA** | 129 | 46,51 | 17,83 | 42,63 | 38,76 |
| **MSc** | 139 | 52,51 | 18,71 | 48,92 | 32,37 |
| **MTech** | 146 | 49,32 | 14,38 | 56,84 | 28,08 |

* **Nhận xét:**

Stress mức trung bình là phổ biến nhất ở cả học sinh lẫn sinh viên → dấu hiệu cảnh báo stress kéo dài.

Stress mức cao xuất hiện nhiều ở bậc đại học, đặc biệt BA và Btech (27% và 25%), cho thấy áp lực học tập và định hướng nghề nghiệp cao sau khi tốt nghiệp.

Tỷ lệ lo âu (Anxiety) cao ở hầu hết các nhóm (46–60%) → Dựa vào chỉ số cho thấy đây là vấn đề phổ biến nhưng cần có chương trình hỗ trợ tinh thần toàn diện.

BSc có tỷ lệ lo âu cao nhất (60%) mặc dù stress cao chỉ ở mức trung bình (16,47%). Điều này cho thấy lo âu và stress không hoàn toàn tương đồng – sinh viên BSc có thể lo âu về thi cử nhưng không đạt mức stress cao .

Nhóm học sinh cấp 2–3 có stress tăng dần theo từng lớp, cao nhất ở Class 11 – giai đoạn chuẩn bị cho kỳ thi lớn. Tuy nhiên, nhìn vào số liệu vẫn thấy mức stress cao nắm ở Class 8 cũng thấy rằng có sự lo lắng khi bước vào Class 8 những năm đầu cho sự định hướng tương lai về con đường chọn khối ngành, nghề nghiệp và trường đại học chuẩn bị theo khi kết thúc Class 12.

Nhóm sau đại học có stress ổn định hơn → khả năng tự điều chỉnh và quản trị stress tốt hơn.

Nhóm Mtech và nhóm Class 9 có tỷ lệ stress cao thấp nhất ( 14,38 và 13,63). Như đã nói ở trên nhóm Mtech là nhóm cao học nên sinh viên đã trưởng thành và điều tiết mọi yếu tố tốt hơn khiến tỷ lệ này giảm xuống. Còn đối với lứa Class 9 là lứa tuổi còn nằm lưng chừng chưa đối mặt áp lực thi đại học và đã qua giai đoạn lo lắng định hướng ở Class 8 nên nhóm này trở nên thoải mái hơn.

### 3.3.2. Phân tích mức độ stress và thời gian ngủ

**Mục tiêu:**  
Khám phá mối quan hệ giữa Stress Level và Sleep Duration để đánh giá vai trò của giấc ngủ đối với sức khỏe tinh thần.

**Aggregation Pipeline:**

**$group:** nhóm sinh viên theo Stress Level.

**$avg:** tính thời gian ngủ trung bình trong mỗi nhóm.

**$median:** tính thời gian ngủ trung vị trong mỗi nhóm.

**$min:** thời gian ngủ tối thiểu của mỗi nhóm

**$max:** thời gian ngủ tối đa của mỗi nhóm

|  |  |
| --- | --- |
| C:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\16B8FF9B.tmp | C:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\D74B7117.tmp |

* **Kết quả phân tích:**

**Unit:**hrs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Stress Level** | **Mean Sleep** | **Median Sleep** | **Min** | **Max** | **Số lượng** |
| Low | 6,55 | 6,60 | 4,00 | 12,00 | 328 |
| High | 6,52 | 6,50 | 4,10 | 12,00 | 181 |
| Medium | 6,41 | 6,40 | 3,00 | 9,00 | 494 |
| Very High | 6,05 | 6,05 | 5,00 | 7,10 | 2 |

* **Nhận xét:**

Không có mối tương quan giữa thời gian ngủ và mức độ Stress. Điều này đi ngược lại với giả thuyết phổ biến rằng nguyên nhân ngủ ít là do stress cao.

Thời gian ngủ trung bình rất đồng đều giữa các nhóm stress (trên 6 tiếng). Chỉ chênh lệch khoảng từ 8 – 40 phút. Điều này chứng tỏ không đơn thuần thiếu ngủ do stress.

Nhóm ngủ nhiều trên 7h vẫn không giảm stress. Tuy nhiên đặt ra một vấn đề khác, nhóm ngủ nhiều vẫn bị stress ở mức cao => ngủ quá nhiều cũng là dấu hiệu mắc bệnh trầm cảm hoặc né tránh ngại giao tiếp.

Tuy nhiên, nhìn chung số liệu vẫn chưa rõ ràng vì nhóm very high stress có số lượng quá ít nên không có ý nghĩa thống kê cao.

=> Từ đó cũng có thể kết luận rằng mức độ stress và thời gian ngủ không có sự tương quan với nhau.

**3.3.3. Giới tính và độ tuổi có tỷ lệ Stress cao nhất**

**Mục tiêu:**  
Xác định nhóm nhân khẩu học có nguy cơ cao dựa trên High Stress Level, Gender và Age.

**Aggregation Pipeline:**

**$bucket:** Chia tuổi thành các nhóm khác nhau: (0-16{≤ 16}, 17-18, 19-21, 22-25, ≥ 26)

**$unwind:** Giải nén các mảng thành từng sinh viên.

**$group** (2 lần):

Lần 1: Nhóm theo (age group, gender, stress).

Lần 2: Nhóm lần nữa theo (age group, gender) để tính % stress cao.

**$project:** Tính tỷ lệ % = Stress High/Tổng sinh viên tương ứng nhóm tuổi và giới tính.

**$sort:** sắp xếp giảm dần theo % stress.

**$limit:** 10 chỉ lấy 10 kết quả đầu tiên.

=> Kết quả chạy Pipeline nằm ở mục sau kết quả phân tích.

**Kết quả sau khi chạy Pipeline:**

|  |  |
| --- | --- |
|  |  |

* **Kết quả phân tích top 10 nhóm (Gender & Age) có High Stress percent cao nhất:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gender** | **Age Group** | **Số lượng đối tượng** | **Số lượng đối tượng High Stress** | **% High Stress** |
| Other | ≤ 16 | 9 | 4 | 44,44% |
| Other | 17-18 | 10 | 3 | 30,00% |
| Other | 19-21 | 14 | 4 | 28,57% |
| Male | ≥26 | 35 | 9 | 25,71% |
| Other | ≥26 | 4 | 1 | 25,00% |
| Female | 19-21 | 133 | 32 | 24.06% |
| Other | 22-25 | 13 | 3 | 23,08% |
| Female | 22-25 | 143 | 28 | 19,58% |
| Male | ≤ 16 | 91 | 17 | 18,68% |
| Male | 19-21 | 111 | 20 | 18,01% |

* **Nhận xét:**

Kết quả phân tích Stress High theo giới tính và độ tuổi cho thấy mức độ căng thẳng có sự biến động đáng kể giữa các nhóm nhân khẩu học. Mức Stress High dao động rất rộng, từ 17% đến hơn 66%, cho thấy các nhóm tuổi – giới có mức chịu áp lực khác nhau rõ rệt.

Trước hết, các nhóm có Stress High trên 50% (60–66.7%) thuộc nhóm có nguy cơ đặc biệt cao. Khi hơn một nửa cá nhân trong một nhóm tuổi–giới tính rơi vào stress nặng, điều này phản ánh sự kết hợp của các yếu tố nhạy cảm như áp lực học tập, giai đoạn chuyển tiếp quan trọng hoặc trách nhiệm xã hội gia tăng. Những nhóm này có thể nằm trong độ tuổi trưởng thành trẻ (18–23), vốn được ghi nhận là giai đoạn căng thẳng nhất do sự thay đổi mạnh về học tập, nghề nghiệp và quan hệ xã hội.

Nhóm Stress High từ 33% đến 42% biểu thị mức stress cao nhưng không nghiêm trọng nhất. Những nhóm này nhiều khả năng đang chịu áp lực vừa phải nhưng kéo dài. Đây là mức dễ chuyển sang stress nặng nếu không có sự hỗ trợ phù hợp.

Đa phần các nhóm tuổi–giới tính rơi vào mức Stress High 20–30%, thể hiện mức stress "chuẩn" thường gặp trong môi trường học thuật hoặc sinh hoạt bình thường. Tuy mức này không quá cao, nhưng kéo dài có thể ảnh hưởng đến giấc ngủ, động lực học tập và khả năng tập trung.

Đáng chú ý, nhóm nữ 22 tuổi có Stress High khoảng 29.7%, phù hợp với xu hướng chung trong nhiều nghiên cứu: nữ giới ở độ tuổi đại học và sau đại học có xu hướng trải nghiệm stress nhiều hơn do tính nhạy cảm cảm xúc, áp lực kỳ vọng xã hội và khối lượng nhiệm vụ lớn.

Nhóm nữ tuổi 19 – 21 tuổi có tỷ lệ stress High khoảng (24.06%). Đây là độ tuổi chuẩn bị tốt nghiệp đại học, tìm việc khác với nam giới thì nữ giới áp lực ngoài vấn đề học tập còn thêm vấn đề xã hội.

Ngược lại, nhóm nam 17 tuổi có Stress High chỉ 17.78%, thuộc nhóm thấp nhất. Điều này phù hợp với đặc trưng: nam giới thường báo cáo stress thấp hơn, và ở tuổi 17 (cuối trung học), mức áp lực chưa cao như nhóm sinh viên đại học.

Nhóm nam Male ≥26 tuổi tỷ lệ High Stress nhìn có vẻ trong top những con số đó cũng không cao. Độ tuổi này đã trưởng thành nên việc điều tiết và kiểm soát công việc đã hình thành. Tuy nhiên, vẫn có một vài trường hợp stress cao có thể do việc gia đình và công việc dồn dập không quản lý kịp. Hoặc nằm trong nhóm muốn thay đổi chuyên ngành nên stress cao.

Nhóm giới tính “Other” luôn có Stress High từ 23% đến 44%, trải đều ở mọi độ tuổi cao hơn so với Female và Male. Nhóm này có thể hiểu rằng nằm ở mặt vấn đề góc nhìn của xã hội đối mặt với phân biệt giới tính, áp lực xã hội, tự cố gắng để học thật tốt có thể có chỗ đứng trong xã hội bù lại cho việc giới tính và thiếu hỗ trợ tâm lý phù hợp.

Nhóm Other ≤ 19 tuổi là nhóm cảnh báo cao (28,57% - 44,44%). Giai đoạn thiếu niên với vấn đề giới tính bản thân, áp lực học tập và gia đình cao hơn các bạn khác.

* **Tổng thể, dữ liệu khẳng định:**

Giới tính: Other có xu hướng Stress High cao hơn nam và nữ.

Độ tuổi: Stress tăng dần từ tuổi thiếu niên → tuổi sinh viên → đầu trưởng thành. Cụ thể hơn là độ tuổi từ 19 – 21 tuổi. Năm trong giai đoạn đại học những năm cuối, quyết định chuyên ngành, thực tập, đồ án và chuẩn bị công việc sau khi ra trường.

Mức chênh lệch rất lớn (17% → 66%) cho thấy các yếu tố cá nhân và xã hội có ảnh hưởng mạnh tới stress.

Những kết quả này giúp xác định các nhóm có nguy cơ cao để ưu tiên triển khai các biện pháp hỗ trợ tâm lý, cải thiện giấc ngủ, quản lý thời gian và tư vấn sức khỏe tinh thần.

Cần phải có chương trình hỗ trợ đặc biệt cho nhóm “Other”: Tư vấn tâm lý với giới tính, nhóm hỗ trợ đồng giới, tạo môi trường thân thiện với những người trong nhóm “Other”.

Can thiệp sớm cho thanh thiếu niên (≤16): Giáo dục sức khỏe tinh thần, những kỹ năng đối phó với áp lực.

Hỗ trợ sinh viên chuẩn bị tốt nghiệp trong độ tuổi (19 – 21 tuổi): Như là tư vấn nghề nghiệp, định hướng, hướng dẫn phỏng vấn, kỹ năng mềm, networking.

### 3.3.4. Nhóm chưa cải thiện học tập (Academic Performance Change = "Declined")

**Mục tiêu:**  
Đánh giá Stress Level, thời gian hoạt động thể chất, và thời gian ngủ của sinh viên có kết quả học tập suy giảm.

**Aggregation Pipeline:**

**🢭 $match**: lọc sinh viên có Academic Performance Change = "Declined".

**🢭 $facet:** Chạy nhiều nhóm truy vấn song song theo nhóm lọc “Declined” như:

**🢭 $group:** Số lượng sinh viên “Declined” theo Stress Level:

**$group:** Trung bình thời gian ngủ, thời gian trên màn hình, thời gian hoạt động thể chất theo “Declined”.

**🢭 $group:** Tách riêng từng mức stress sinh viên “Declined”, tính trung bình của từng chỉ số trên từng nhóm stress.

**Kết quả sau khi chạy Pipeline:**

|  |  |
| --- | --- |
| C:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\B93523D.tmp | C:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\E409C53.tmp |
| C:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\48166CD9.tmp | |

* **Kết quả phân tích:**

Tổng số sinh viên Declined: 299 (29,75%)

**Tổng quan các chỉ số theo nhóm Declined:**

|  |  |
| --- | --- |
| Trung bình thời gian màn hình (hrs/ngày) | 6,68 |
| Trung bình thời gian ngủ (hrs/ngày) | 6,45 |
| Trung bình thời gian hoạt động (hrs/tuần) | 5,37 |

**Tổng quan Stress Level theo nhóm Declined:**

|  |  |  |
| --- | --- | --- |
| **Stress Level** | **Số lượng** | **Tỷ lệ (%)** |
| **Low** | 90 | 30,10% |
| **Medium** | 157 | 52,51% |
| **High** | 51 | 17,06% |
| **Very High** | 1 | 0,33% |

**So Sánh các chỉ số sinh hoạt: Declined và thông tin chung:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Chỉ số** | **Trung bình Declined** | **Tổng quan trung bình** | **Chênh Lệch** |
| **Screen Time (hrs/ngày)** | 6,68 | 6,93 | -0,25 |
| **Sleep Duration (hrs/ngày)** | 6,45 | 6,47 | -0,03 |
| **Physical Activity (hrs/tuần)** | 5,37 | 5,22 | +0,15 |

* **Nhận xét:**

Nhóm Declined có tỷ lệ Stress Medium cao (52,51%). Điều này cho thấy stress vừa phải có thể ảnh hưởng tiêu cực đến học tập nếu không điều tiết tốt mọi thứ xung quanh.

Thời gian xem màn hình thấp hơn so với thông tin chung (-0,25h) Điều này không đúng với việc dành thời gian học trên màn hình nhiều sẽ đạt hiệu quả hoặc có thể động lực học tập bị thiếu.

Nhìn chung các chỉ số sinh hoạt nhóm Declined gần như không có khác biệt nhiều so với tổng quan chung. Điều này chứng minh rằng giảm sút học tập không phải do lối sống kém mà có thể do phương pháp học tập, bài giảng, động lực, yếu tố tâm lý hoặc do môi trường xung quanh đang ngồi học.

Tuy nhiên, Nhóm Declined có High stress chiếm đến 17% - Đây là nhóm cần can thiệp khẩn cấp vì kết hợp yếu tố stress và học lực giảm => Tinh thần sẽ không ổn định, dẫn đến các bệnh tâm lý.

* **Giải Pháp Đề Xuất:**

Không nên chỉ tập trung vào lối sống: Các chỉ số sinh hoạt không phải nguyên nhân chính.

Tăng cường kỹ năng học tập: Study skills workshop, time management, note-taking, active learning.

Động lực và mục tiêu: Career counseling, động lực học tập, kết nối bạn bè học nhóm.

Can thiệp tâm lý cho nhóm Medium Stress: Giúp họ không tiến triển thành High Stress.

Nhóm chưa cải thiện trong học tập có đặc trưng: ngủ thiếu – vận động thấp – stress cao.

Những yếu tố này kết hợp với nhau tạo thành những rào cản lớn đối với sự tiến bộ trong học tập. Nhóm này nên được ưu tiên trong các chương trình can thiệp, tập trung vào:

cải thiện chất lượng giấc ngủ,

tăng mức độ vận động hàng tuần,

giảm stress học tập thông qua tư vấn tâm lý hoặc điều chỉnh lịch học.

### 3.3.5. Nhóm cải thiện học tập (Academic Performance Change = "Improved")

**Mục tiêu:**  
Đánh giá tác động của Stress Level lên Screen Time, Sleep Duration và Physical Activity của nhóm cải thiện học tập.

**Aggregation Pipeline:**

**🢭 $match:** lọc sinh viên Academic Performance Change = "Improved".

**🢭 $group:** tính trung bình ScreenTime, SleepDuration và PhysicalActivity theo StressLevel.

|  |
| --- |
|  |

* **Kết quả phân tích:**

Tổng số sinh viên Improved: 54(5,37%)

|  |  |  |  |
| --- | --- | --- | --- |
| **Chỉ số** | **High stress + Improved** | **Tổng quan trung bình** | **Chênh Lệch** |
| **Screen Time (hrs/ngày)** | 7,09 | 6,93 | +0,16 |
| **Sleep Duration (hrs/ngày)** | 6,23 | 6,47 | -0,24 |
| **Physical Activity (hrs/tuần)** | 5,03 | 5,22 | -0,19 |

* **Nhận xét:**

**Stress cao → Screen Time cao (7+ giờ/ngày)**, phù hợp với xu hướng chung, là yếu tố rủi ro về sức khỏe tinh thần .

Dù stress lớn và Screen Time cao, nhóm này **vẫn cải thiện học tập**, có thể do dùng nhiều thời gian màn hình vào hoạt động học.

Tuy nhiên, Screen Time cao khiến giấc ngủ giảm (6.23h) và vận động thấp (5h/tuần).

Nhóm này *có kết quả học tập tốt nhưng phải đánh đổi sức khỏe*: căng thẳng – ngủ kém – ít vận động.

* **Giải Pháp Đề Xuất:**

**Giảm screen time xuống ≤ 5 giờ/ngày**

Tắt thông báo khi học

Không dùng điện thoại 1 giờ trước khi ngủ

**Tăng giấc ngủ lên 7–8 giờ/đêm**

Ngủ và dậy giờ cố định

Không học quá khuya kéo dài

**Tăng nhẹ hoạt động thể chất (5 → 6–7 giờ/tuần)**

Đi bộ nhanh 30 phút/ngày

Tham gia thể thao nhóm

### 3.3.6. Nhóm nguy cơ cực cao

**Mục tiêu:**  
Xác định sinh viên có nguy cơ cực cao dựa trên nhiều tiêu chí: Stress cao, Anxious = Yes, Screen Time vượt ngưỡng, Sleep thấp.

**Điều kiện:**

Stress Level = High

Anxious Before Exam = "Yes"

Screen Time > 6 giờ/ngày

Sleep Duration < 6 giờ/ngày

Physical Activity  <2 giờ/tuần (low)

**Aggregation Pipeline:**

**🢭 $match**: lọc theo các điều kiện trên.

**🢭 $facet**: chạy nhiều truy vấn tổng hợp song song trên cùng một bộ lọc:

**🢭 Count:** đếm tổng số sinh viên thuộc nhóm nguy cơ cao.

**Gender:** Liệt nhóm nguy cơ cao theo giới tính.

**Performance:** Liệt kê kết quả học tập theo 3 nhóm (Improved, Same, Declined).

**Metrics:** Tính giá trị trung bình cho các chỉ số.

**Sau khi chạy Pipeline ta được kết quả**

|  |  |  |
| --- | --- | --- |
| **C:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\F47D3D4F.tmp** | **C:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\EB174335.tmp** | **C:\Users\Sam Py\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\533D00B.tmp** |

* **Kết quả phân tích:**

Tổng số sinh viên nguy cơ cực cao: 8 (0,79%). Ở đây đặt theo các điều kiện và yếu tố đặt ra. Có thể thay đổi nếu có sự biến đổi trong nghiên cứu sức khỏe và lối sống.

Phân bố theo giới tính:

* Female: 7
* Male: 1
* Phân bố theo Academic Performance Change:
* Improved: 4 (50%) => Điểm đặc biệt
* Same: 2 (25%)
* Declined: 2 (25%)

Thống kê chỉ của nhóm nguy cơ cao:

|  |  |  |  |
| --- | --- | --- | --- |
| **Chỉ số** | **Trung bình High Risk** | **Tổng quan trung bình** | **Chênh Lệch** |
| Screen Time (hrs/ngày) | 10,45 | 6,93 | +3,32 |
| Sleep Duration (hrs/ngày) | 4,61 | 6,47 | -1,86 |
| Physical Activity (hrs/tuần) | 1,03 | 5,22 | -4,19 |

* **Nhận xét:**

Screen Time quá cao (10+ giờ/ngày) → tăng nguy cơ căng thẳng, mỏi mắt, giảm tập trung, giảm giấc ngủ.

Ngủ quá ít (4.6 giờ) → ảnh hưởng trực tiếp tới khả năng học tập dài hạn, stress tăng, giảm sức đề kháng.

Hoạt động thể chất thấp (1h/tuần) → chưa đủ để giảm stress, cải thiện tinh thần, tăng năng lượng.

Hiệu quả học tập vẫn “Improved” nhưng có thể không bền vững nếu lối sống này kéo dài.

* **Giải Pháp Đề Xuất:**

**Giảm Screen Time**

Giới hạn **4–6 giờ/ngày**, ưu tiên thời gian học tập trực tuyến có mục tiêu.

Ngắt quãng mỗi 45–60 phút khi học trên máy tính để giảm mỏi mắt.

**Tăng giấc ngủ**

Cố gắng ngủ **7–9 giờ/đêm**.

Thiết lập giờ đi ngủ và dậy cố định.

Giảm dùng điện thoại / thiết bị trước khi ngủ để cải thiện chất lượng giấc ngủ.

**Tăng hoạt động thể chất**

Đạt **5–7 giờ/tuần** hoạt động vừa/khỏe (chạy bộ, đi bộ, thể dục nhịp điệu).

Hoạt động thể chất giúp **giảm stress**, **cải thiện tinh thần**, hỗ trợ học tập hiệu quả hơn.

**Quản lý stress**

Thực hành kỹ thuật giảm căng thẳng: thiền, yoga, hít thở sâu.

Phân bổ thời gian học hợp lý, ưu tiên nhiệm vụ quan trọng, tránh học quá dài liên tục.

**Theo dõi và hỗ trợ dài hạn**

Nhóm này cần **giám sát liên tục** lối sống và stress.

Tư vấn học thuật hoặc tâm lý có thể giúp duy trì hiệu quả học tập mà vẫn bảo vệ sức khỏe.

## **3.4. Tối ưu hiệu năng bằng Index trong MongoDB**

Để tăng tốc độ truy vấn trong các pipeline trên, các trường quan trọng cần tạo Index:

* **Các Index đề xuất:**

EducationLevel

StressLevel

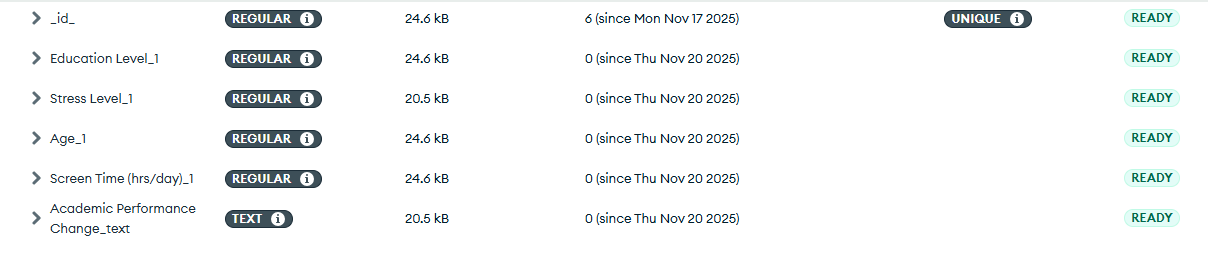
AnxiousBeforeExam

Gender

Age

ScreenTime

AcademicPerformanceChange



**- Ý nghĩa:**  
Index giúp giảm thời gian quét toàn bộ dữ liệu, nâng cao hiệu năng khi thực hiện nhiều pipeline với điều kiện $match phức tạp. Việc tối ưu này đảm bảo phân tích dữ liệu nhanh và hiệu quả.

## **3.5. Kết luận chương 3**

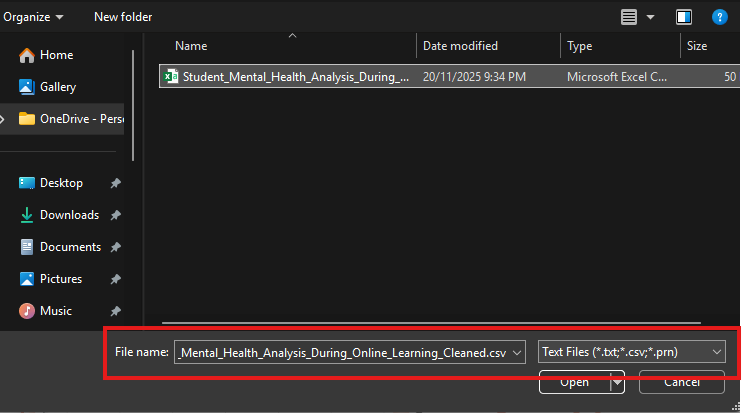
Trong Chương 3, nghiên cứu đã tiến hành phân tích dữ liệu nhằm giải đáp sáu câu hỏi trọng tâm liên quan đến sức khỏe tinh thần, hành vi học tập và các yếu tố ảnh hưởng đến sinh viên. Thông qua hệ thống các pipeline phân tích trên MongoDB, chương này đã phác họa toàn diện những mối quan hệ quan trọng giữa stress, giấc ngủ, thời gian sử dụng thiết bị, mức độ lo âu, trình độ học vấn và sự thay đổi kết quả học tập.

# CHƯƠNG 4 : TRỰC QUAN HÓA DỮ LIỆU

Từ bộ dữ liệu đã được làm sạch và xuất ra ở Chương 2, tiến hành import vào Power BI theo các bước sau:

Bước 1**:** Mở Power BI Desktop.  
 Bước 2: Trên thanh Home, chọn Get data → Text/CSV.  
 Bước 3: Chọn tệp Student Mental Health Analysis During Online Learning.csv.  
 Bước 4: Nhấn Open để xem trước dữ liệu.  
 Bước 5**:** Kiểm tra các lại các cột để đảm bảo dữ liệu nhận diện đúng

Bước 6: Nhấn Transform Data để mở Power Query Editor.  
 Bước 7: Trong Power Query Editor, kiểm tra và thiết lập đúng kiểu dữ liệu cho các cột.  
 Bước 8: Sau khi hoàn tất xử lý dữ liệu, nhấn Close & Apply để tải dữ liệu vào Power BI.  
 Bước 9: Chờ Power BI nạp dữ liệu hoàn chỉnh và sẵn sàng cho bước xây dựng mô hình và trực quan hóa.



## **4.1. Mục tiêu của phần phân tích trên Power BI**

Mục tiêu của báo cáo Power BI là trực quan hóa dữ liệu sinh viên nhằm phát hiện các hành vi có nguy cơ dẫn đến căng thẳng học đường. Báo cáo tập trung vào các KPI chính gồm:

Đánh giá mối liên hệ giữa lối sống (ngủ, hoạt động thể chất, thời gian dùng màn hình) và sức khỏe tâm lý.

Xác định xem **y**ếu tố hành vi có ảnh hưởng đến hiệu suất học tập, Stress hay không.

Giúp phát hiện các yếu tố rủi ro (stress cao, ngủ ít, lo âu trước kỳ thi…).

Hỗ trợ xây dựng các chương trình tư vấn, hỗ trợ sức khỏe tinh thần cho học sinh/sinh viên.

## **4.2. Mô tả cấu trúc và thiết kế Dashboard**

Cấu trúc Dashboard gồm 4 trang chính, với mục tiêu phân tích toàn diện tình trạng sức khỏe tinh thần, hành vi và kết quả học tập của học sinh/sinh viên:

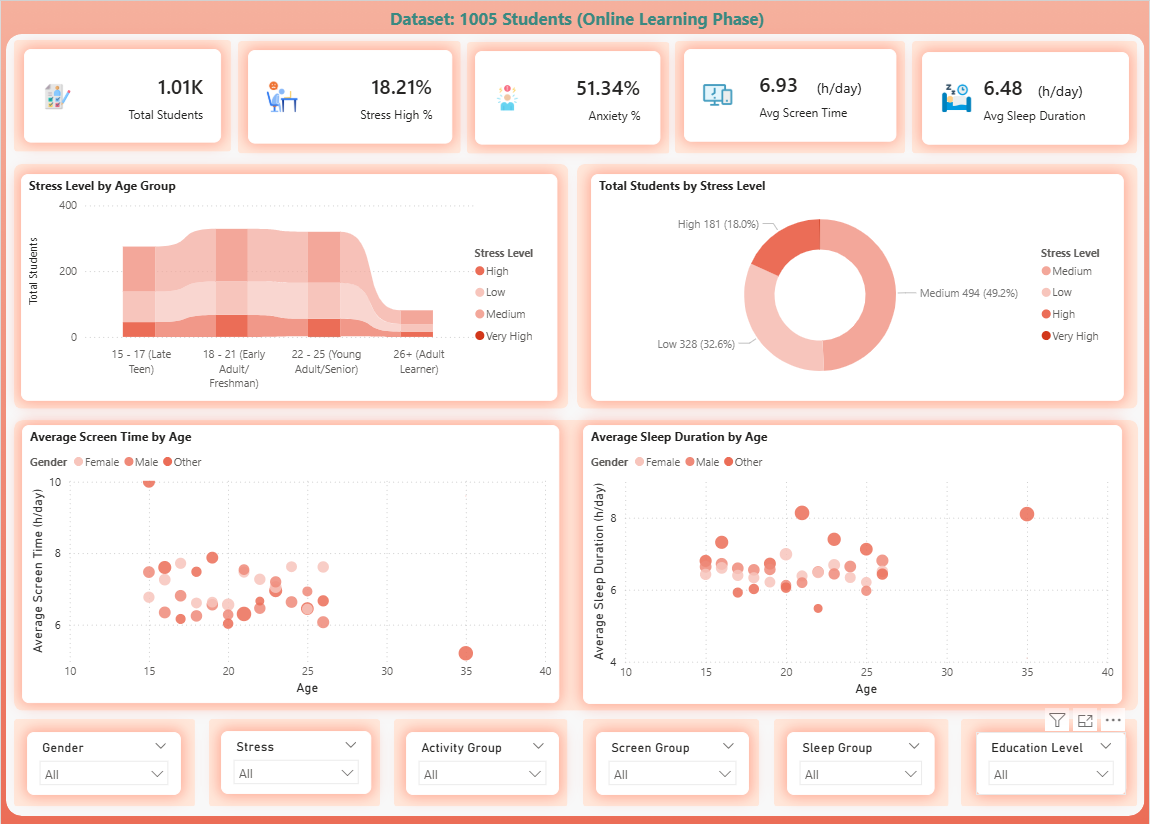
**Page 1: Student Profile & Mental Health Status**

Trang này cung cấp cái nhìn tổng quan về đặc điểm cá nhân và tình trạng sức khỏe tinh thần hiện tại của người học. Nội dung nhằm:

Đánh giá mức độ stress, lo âu và các vấn đề giấc ngủ.

Xác định các nhóm đối tượng có nguy cơ cao hoặc dễ bị tổn thương.

Tạo cơ sở ban đầu để nhận diện nhóm cần được hỗ trợ hoặc theo dõi.



**Page 2: Behavior Impact Analysis**

Trang này phân tích mức độ ảnh hưởng của các hành vi thường ngày đến sức khỏe tinh thần. Một số mối liên hệ thường gặp được làm rõ như:

Ngủ không đủ giấc dẫn đến mức độ stress cao.

Thiếu hoạt động thể chất làm tăng khả năng lo âu.

Thời gian sử dụng thiết bị điện tử kéo dài gây ngủ muộn và giảm khả năng tập trung.

Mục tiêu là xác định các hành vi tiêu cực và mức độ tác động của chúng.



**Page 3: Academic Performance**

Trang này đánh giá ảnh hưởng của các yếu tố tâm lý và hành vi đến kết quả học tập. Phân tích cho phép nhận diện:

Các yếu tố giúp nâng cao thành tích (như ngủ đủ giấc, mức stress thấp).

Các yếu tố làm suy giảm hiệu suất học tập (như stress cao, ngủ ít, dùng màn hình quá nhiều).

Nhờ đó, trang này cho thấy mối liên kết giữa tình trạng sức khỏe tinh thần – hành vi – kết quả học tập.

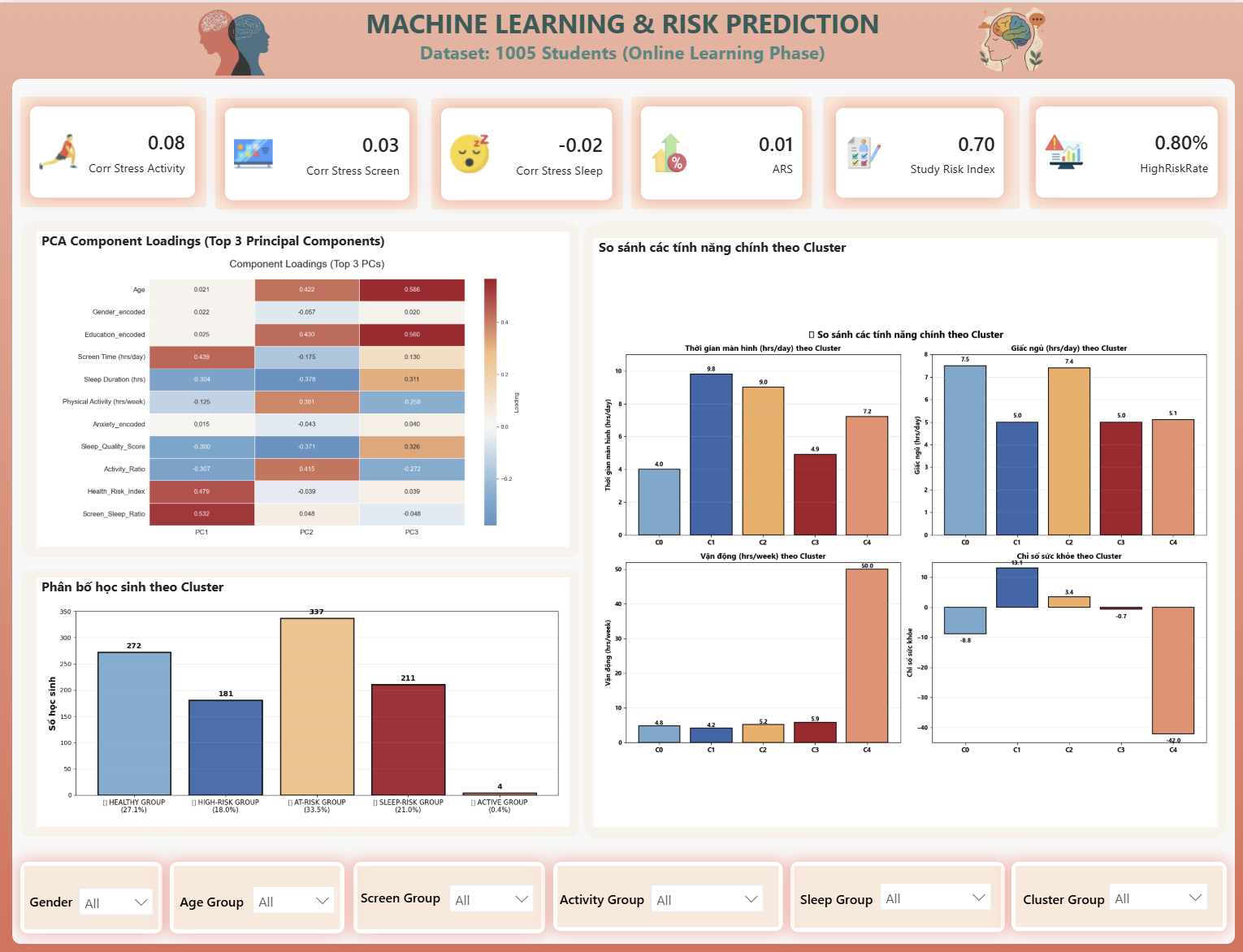
Phát hiện các nhóm học sinh/sinh viên có nguy cơ tiềm ẩn cao. Nhóm có thể vượt qua những áp lực về stress để đạt được kết quả tốt nhưng có nguy cơ ảnh hưởng không tốt đến sức khỏe.



**Page 4: ML & Risk prediction**

Dự đoán Rủi ro và Học máy (Machine Learning & Risk Prediction), được xây dựng dựa trên tập dữ liệu của 1005 học sinh trong giai đoạn học trực tuyến (Online Learning Phase).

Trang này cung cấp một cái nhìn toàn diện về cách các yếu tố như thời gian ngủ, vận động, và thời gian sử dụng màn hình ảnh hưởng đến việc phân loại rủi ro sức khỏe/học tập của học sinh trong giai đoạn học trực tuyến.



## **4.3. Insight chính (Key Findings)**

### 4.3.1. Nhóm sinh viên lớn tuổi (~35 tuổi) là ngoại lệ với thời gian ngủ cao và thời gian màn hình thấp

**- Dữ liệu minh chứng:**

Trong biểu đồ **“Average Sleep Duration by Age”**, điểm dữ liệu ở tuổi **khoảng 35** thể hiện mức **ngủ trung bình cao nhất (~9 giờ/ngày)**.

Ở biểu đồ **“Average Screen Time by Age”**, điểm dữ liệu cùng tuổi này có **screen time thấp nhất**, dưới **6 giờ/ngày**, thấp hơn đáng kể so với phần lớn sinh viên trẻ (thường 6–9 giờ/ngày).

* **Phân tích:**

Sinh viên ở độ tuổi trưởng thành (~35 tuổi) có xu hướng **ngủ đủ giấc và kiểm soát thời gian sử dụng thiết bị tốt hơn**. Điều này cho thấy họ đã hình thành kỹ năng quản lý thời gian hiệu quả, có sự cân bằng rõ rệt giữa công việc – học tập – nghỉ ngơi. Họ cũng ít bị ảnh hưởng bởi các thói quen tiêu cực như dành quá nhiều giờ trước màn hình, vốn phổ biến ở sinh viên trẻ tuổi. Đây là nhóm ngoại lệ đáng chú ý trong dataset, gợi ý giá trị tham chiếu về thói quen sinh hoạt lành mạnh.

### 4.3.2. Giới tính “other” độ tuổi ảnh hưởng đến stress

**- Dữ liệu minh chứng:**

Trong biểu đồ Stress level by age group ta có thể thấy Gender other có sự thay đổi theo từng giai đoạn:

Từ 15 - 17, 22 - 25 và 26+ tuổi: Thì đối tượng này số lượng stress cao rất ít

Từ 18 - 21 tuổi: Những đối tượng này đang trong giai đoạn độ tuổi quyết định rất nhiều thứ tương lai, học tập, kỳ vọng gia đình và chính bản thân phải chọn con đường nào. Ngoài ra, về vấn đề tâm lý chung là những đối tượng này thường hay tư ti rất e dè vì định kiến xã hội và tâm lý phải học thật tốt để có tương lai tốt hơn bù cho việc giới tính. Thông qua dữ liệu thời gian học trên màn hình của những đối tượng này thời gian trên 6 giờ/ngày thấp hơn so với hai giới tính còn lại.

Nhìn vào biểu đồ tỷ lệ phần trăm học sinh giới tính other bị stress level ta có thể thấy thêm là đối tượng này phân bổ rất điều (30,40,30), không chênh lệch nhiều như ở hai giới tính còn lại. Đáng lo ngại nhất là phần high stress và medium chiêm 30 - 40.

* **Phân tích:**

Học sinh từ 18-21 giới tính other mang áp lực nhiều hơn so với hai giới tính kia. Điều này cho thấy ngoài những chỉ số về thời gian sức khỏe, độ lo lắng thì những bạn này đang bị một vấn đề khác mang tên tâm lý xã hội. Những tâm lý đó là những tác nhân bên ngoài ảnh hưởng mạnh mẽ nhất đến vấn đề stress của các bạn như: sự khác biệt, ánh nhìn soi mói, kỳ thị, kỳ vọng bản thân để học thật tốt để vượt qua**.**

### 4.3.3. Cả thời gian sử dụng màn hình và hoạt động thể chất đều tăng cùng với mức độ Stress.

**- Dữ liệu (Biểu đồ "How Activity, Screen Time, and Sleep Relate to Stress"):**

Thời gian sử dụng màn hình trung bình (**Avg Screen**) tăng từ **6.5 h/ngày** (Stress Low) lên **7.2 h/ngày** (Stress Very High).

Hoạt động thể chất trung bình (**Avg Activity**) cũng tăng từ **6.9 h/tuần** (Stress Low) lên **7.4 h/tuần** (Stress Very High).

* **Phân tích:**

**Tương tác Màn hình:** Sinh viên có mức độ stress cao hơn đang dành nhiều thời gian hơn trước màn hình (có thể là học tập cường độ cao hoặc sử dụng màn hình như một cơ chế đối phó/xao nhãng).

**Hoạt động Thể chất:** Việc tăng hoạt động thể chất ở nhóm stress cao có thể là nỗ lực tự điều chỉnh/tìm kiếm lối thoát cho sự căng thẳng, những hành vi này có vẻ chưa đủ để làm giảm mức độ stress.

### 4.3.4. Hoạt động thể chất càng cao thì Stress Level càng tăng — cho thấy sinh viên stress nặng tập thể dục nhiều hơn (mang tính đối phó)

**- Dữ liệu minh chứng**

Ở biểu đồ **“Stress Level by Avg Physical Activity”**, Physical Activity tăng theo mức độ stress:

**Low Stress:** ~4.8 giờ/tuần

**Medium Stress:** ~5.3 giờ/tuần

**High Stress:** ~5.7 giờ/tuần

**Very High Stress:** ~6.9 giờ/tuần → cao nhất

* **Phân tích**

Điều thú vị là sinh viên có Stress Very High lại có thời gian hoạt động thể chất cao nhất. Điều này gợi ý rằng hoạt động thể chất có thể đang được sử dụng như một **cơ chế đối phó** (coping mechanism) khi stress tăng, hoặc do họ cố gắng cải thiện sức khỏe tinh thần nhưng chưa đủ để bù lại các yếu tố tiêu cực khác như thiếu ngủ hoặc Screen Time cao. Kết quả này chỉ ra rằng chỉ tăng cường vận động là **chưa đủ** để giảm stress nếu thói quen ngủ và sử dụng màn hình không được cải thiện đồng thời.

### 4.3.5. Khả năng vượt qua các tác động ảnh hưởng đến tinh thần – Phân tích Level 2

Khả năng phục hồi tinh thần (Academic Resilience) của nhóm Level 2 thể hiện rõ rệt qua các chỉ số hành vi học tập và sức khỏe tinh thần. Nhóm này cho thấy sự ổn định cao hơn so với các mức còn lại, dù phải đối mặt với áp lực học tập và thời gian học trực tuyến kéo dài.

* **Dữ liệu (Biểu đồ “Stress Numeric Level – Sleep – Screen Time – Activity”)**

**Số lượng sinh viên Level 2 chiếm tỷ lệ lớn nhất trong mẫu:**  
– Level 2: **494 sinh viên**, cao nhất trong tất cả các nhóm.

**Nhóm Level 2 có khả năng thích nghi tốt với các yếu tố gây stress:**  
– Screen Time chủ yếu ở mức **Medium (4–8 giờ/ngày)**  
– Giấc ngủ nằm ở hai nhóm chính:  
• Ngủ ít (<6 giờ): **205 sinh viên**  
• Ngủ trung bình (6–8 giờ): **203 sinh viên**  
– Hoạt động thể chất phần lớn thuộc nhóm **Short (<6 giờ/tuần)**

**- Phân tích:**

**a. Khả năng duy trì hiệu suất dù chịu áp lực**

Sinh viên Level 2 không thuộc nhóm stress thấp nhất, nhưng vẫn giữ được sự ổn định về tinh thần và học tập. Điều này phản ánh khả năng:

Tập trung cao,

Tự điều chỉnh cảm xúc,

Duy trì động lực học tập trong môi trường trực tuyến cường độ cao.

→ **Áp lực không làm suy giảm hiệu suất mà trở thành động lực kích hoạt năng lực học tập.**

**b. Khả năng phục hồi tinh thần (ARS) vượt trội**

Chỉ số ARS của nhóm Level 2 nằm ở mức **từ trung bình đến cao**, cao hơn mức ARS chung của toàn bộ dataset. Điều này cho thấy:

Họ thích nghi tốt trước các tác động từ việc học trực tuyến,

Screen Time cao hoặc giấc ngủ giảm không gây suy giảm mạnh như ở nhóm khác,

Nhóm có xu hướng cân bằng được giữa nỗ lực học tập và trạng thái tinh thần.

→ **Level 2 là nhóm “high-performer”, sở hữu năng lực chống chịu tốt với stress.**

### 4.3.6. Mô hình và con số báo động ở phân cụm nhóm At-Risk (33,53%) cần hỗ trợ thực thi

Dựa trên hình phân cụm tổng số học sinh theo 5 nhóm, ta có thể thấy được nhóm nguy cơ tiềm ẩn về vấn đề sức khỏe khi học online đang chiếm một số rất nổi bật 337 học sinh/sinh viên tương đương 33,53%. Do chỉ chiếm có 1/3 dữ liệu nên có thể bị bỏ qua. Tuy nhiên, những nhóm này đang ở trạng thái “chuyển tiếp” mặc dù hành vi hoạt động chữ “xấu” nhưng đang dần trở nên xấu hơn gia nhập nhóm High-Risk với số lượng 181 học sinh/sinh viên tương đương với 18,01%. Con số này nếu chúng ta không tìm hiểu và can thiệp kịp thời thì nhóm High-risk sẽ chiếm trên 50% dữ liệu một con số đáng báo động.

**- Nguyên nhân sẽ rất nhiều lý do khác nhau:**

Đầu tiên là các yếu tố tương quan nó là một cái bẫy vì nhìn vào biểu đồ PCA và biểu đồ phân cụm theo hành vi, thì chúng ta thấy được phần lớn khác biệt trong dữ liệu không đền từ hành vi và từ đặc tính nhân khẩu học: Tuổi, trình độ học vấn, giới tính. Các hành vi như giấc ngủ, xem màn hình, hoạt động có tác động nhưng nằm mở mức yếu chưa rõ ràng. Tuy nhiên, cũng góp một phần vào đánh giá được các mẫu (pattern) rủi ro khác nhau trong lối sống.

Thứ hai nhìn lại các biểu đồ ta có thể thấy được Stress là tập hợp không phải tự một yếu tố, mà là tổ hợp từ nhiều yếu tố tương tác lẫn nhau: Áp lực học tập, kỳ vọng gia đình, khả năng quản lý cảm xúc, tuổi tác, giới tính (đa phần từ áp lực những định kiến của xã hội), trình độ học vấn (độ khó của bài tập). Từ đó thấy được ngoài các yếu tố hành vi hoạt động thì sẽ phải tính thêm các hành vi khác nhau tâm lý lo lắng, sợ sệt, ngại tiếp xúc giao tiếp người khác.

**- Giải pháp:**

Nhóm At-Risk là nhóm cần phải can thiệp kịp thời. Chuyển nhóm At-Risk => Healthy càng sớm càng tốt. Giảm nhóm High-Risk. Tăng tỷ lệ Improved song song đó giám sát tâm lý những bạn Improved tránh các bạn sử dụng biện pháp tiêu cực ngắn hạn dẫn đến suy nhược.

Xây dựng hệ thống cảnh báo sớm thông qua mô hình:

Screen Time + Sleep + Activity + Test Scores

↓

Mô hình dự đoán cluster

↓

Output: High-Risk, At-Risk, Sleep-Risk, Healthy

↓

Phụ huynh, giáo viên, tư vấn viên

Tuy các biến hành vi không tác động mạnh nhưng nó là bước đầu trong việc phát hiện những rủi ro rõ ràng và dễ dàng nhất. Vì tâm lý con người thường hay che giấu những biểu hiện bên trong.

Tạo ra các buổi Workshop: "Không ai hoàn hảo". Tư vấn cân bằng lại cuộc sống healthy thông qua quản lý kỳ vọng cá nhân, tâm lý cởi mở, không thay đổi hành vi mà cải thiện hành vi tốt hơn so với thể chất mỗi cá nhân.

## **4.4. Kể chuyện dữ liệu (Data Storytelling)**

Bức tranh về sức khỏe tinh thần của sinh viên trở nên rõ ràng hơn khi ta lần theo những câu chuyện mà dữ liệu đang âm thầm kể lại. Tất cả bắt đầu từ sự khác biệt giữa các nhóm tuổi – một điểm nhấn đầy bất ngờ. Khi phần lớn sinh viên trẻ phải vật lộn với thiếu ngủ, phụ thuộc vào màn hình, và gia tăng stress, thì đâu đó trong dataset lại xuất hiện một “điểm sáng” hiếm hoi: nhóm sinh viên khoảng 35 tuổi. Họ ngủ nhiều hơn, dùng màn hình ít hơn và dường như đã tìm thấy nhịp sống cân bằng mà những sinh viên trẻ vẫn đang tìm kiếm. Sự ổn định này gợi ra một chuẩn tham chiếu – rằng cách sống lành mạnh hoàn toàn có thể đạt được, chỉ là ở giai đoạn tuổi trẻ, nhiều yếu tố khiến sinh viên khó duy trì được điều đó.

Tuy nhiên, không phải nhóm tuổi nào cũng đang bước đi trong vùng an toàn ấy. Khi dữ liệu chuyển sang phân tích theo giới tính, một câu chuyện đầy chiều sâu xuất hiện – đó là hành trình áp lực của nhóm giới tính “other”, đặc biệt trong giai đoạn 18–21 tuổi. Trong khi các nhóm tuổi khác của giới tính này có mức stress khá thấp, thì nhóm 18–21 lại chứng kiến mức tăng vọt. Đây là giai đoạn của những lựa chọn lớn đời người: ngành học, tương lai nghề nghiệp, kỳ vọng gia đình. Và với những bạn mang bản dạng giới khác biệt, căng thẳng này không chỉ đến từ học tập mà còn từ ánh nhìn xã hội, từ cảm giác mình phải nỗ lực gấp đôi để chứng minh giá trị. Dữ liệu cho thấy họ dùng màn hình ít hơn – không phải vì thói quen tốt, mà có thể vì sự thu mình, ngại tương tác và cảm giác bị quan sát.

Khi tiếp tục nhìn sâu vào mối liên hệ giữa stress, thời gian dùng màn hình và hoạt động thể chất, dữ liệu bắt đầu kể một câu chuyện nghịch lý. Ta thường tin rằng stress cao thì vận động ít, hoặc ngủ nhiều sẽ giảm stress. Nhưng không: sinh viên càng stress lại càng *vận động nhiều hơn* và *dùng màn hình nhiều hơn*. Ở mức stress Very High, thời gian tập thể dục tăng lên cao nhất, nhưng tinh thần vẫn không cải thiện. Hoạt động thể chất – vốn được xem là giải pháp tốt – nay lại phản ánh một cơ chế đối phó: càng căng thẳng, sinh viên càng cố gắng chạy bộ, tập luyện để giải tỏa, nhưng những nỗ lực này dường như chỉ làm giảm áp lực tạm thời. Bởi gốc rễ của stress không nằm ở việc ít vận động, mà xuất phát từ sự kết hợp của thiếu ngủ, học tập cường độ cao và áp lực xã hội.

Giữa bức tranh đầy nhiễu động đó, một nhóm khác lại nổi lên như “cột trụ” của sự ổn định: nhóm Stress Level 2. Đây là nhóm đông nhất trong dataset – và cũng là nhóm thể hiện khả năng thích nghi tốt nhất. Dù không phải nhóm stress thấp, họ vẫn giữ được hiệu suất học tập, duy trì các thói quen sinh hoạt vừa phải và không bị kéo vào vòng xoáy của stress cao. Screen time nằm trong mức trung bình, giấc ngủ không quá lý tưởng nhưng không bị thiếu trầm trọng, và hoạt động thể chất ở mức đủ dùng. Đặc biệt, chỉ số ARS của nhóm này cao hơn mức chung, cho thấy một khả năng phục hồi tinh thần đáng kể. Nếu nhóm sinh viên 35 tuổi là hình mẫu về thói quen sinh hoạt, thì nhóm Level 2 chính là hình mẫu về *năng lực đối phó với căng thẳng*.

Khi kết nối toàn bộ câu chuyện lại, một thông điệp rõ ràng hiện lên: **Stress không phải chỉ là kết quả của mỗi hành vi, mà là sự giao thoa giữa hành vi, thói quen, tuổi tác và cả yếu tố bên ngoài như xã hội, tâm lý.**

Có những nhóm chịu ảnh hưởng đặc biệt nặng – như giới tính “other” ở tuổi 18–21. Có những nhóm thể hiện được mô hình đối phó hiệu quả – như Level 2. Và cũng có những nhóm mang đến bài học về lối sống cân bằng – như sinh viên lớn tuổi.

Dữ liệu không chỉ cung cấp con số; nó đang kể về cách sinh viên sống, cảm nhận và đối mặt với thế giới xung quanh. Nó chỉ ra những nơi đang “nóng lên”, những nơi cần hỗ trợ, và cả những điểm sáng để dựa vào. Chính những câu chuyện đằng sau con số này sẽ là chìa khóa cho mọi chiến lược hỗ trợ sức khỏe tinh thần trong môi trường học đường.

## **4.5. Kết luận & Khuyến nghị**

* **Kết Luận:**

Dữ liệu cho thấy sức khỏe tinh thần của sinh viên chịu tác động đồng thời từ thói quen sinh hoạt, đặc điểm tuổi tác và yếu tố xã hội.

Nhóm sinh viên lớn tuổi (~35 tuổi) thể hiện lối sống cân bằng: ngủ nhiều, screen time thấp – trở thành chuẩn tham chiếu về thói quen lành mạnh.

Giới tính “other”, đặc biệt giai đoạn 18–21 tuổi, chịu stress cao do áp lực xã hội và kỳ vọng cá nhân, không chỉ từ học tập.

Nhóm stress cao có xu hướng tăng cả screen time và vận động, phản ánh cơ chế đối phó tạm thời thay vì giải pháp bền vững.

Nhóm Stress Level 2 giữ được sự ổn định tốt, vừa chịu áp lực vừa duy trì hiệu suất – là nhóm mẫu có khả năng phục hồi tinh thần mạnh nhất.

Tổng thể, stress của sinh viên mang tính đa chiều, không thể giải quyết bằng một yếu tố đơn lẻ.

* **Kiến nghị:**

**Hỗ trợ tâm lý – xã hội cho sinh viên giới tính “Other”**

Cung cấp dịch vụ tư vấn tâm lý cá nhân và nhóm dành riêng cho sinh viên “Other”, đặc biệt nhóm tuổi 18–21.

Tạo môi trường học tập an toàn, giảm định kiến và tăng cường sự chấp nhận đa dạng giới.

**Cải thiện thói quen sinh hoạt cho sinh viên trẻ**

Triển khai các chương trình hướng dẫn quản lý giấc ngủ và thời gian sử dụng màn hình.

Khuyến khích áp dụng mô hình sinh hoạt lành mạnh từ nhóm sinh viên lớn tuổi (~35 tuổi).

**Can thiệp toàn diện cho nhóm stress cao**

Kết hợp quản lý giấc ngủ, kiểm soát screen time, rèn luyện hoạt động thể chất và kỹ năng đối phó stress.

Tổ chức các workshop về mindfulness, quản lý thời gian và kỹ năng phục hồi tinh thần.

**Nhân rộng mô hình nhóm Stress Level 2**

Nghiên cứu chiến lược tự điều chỉnh cảm xúc, duy trì động lực và cân bằng học tập của nhóm này.

Sử dụng nhóm Level 2 làm mô hình tham khảo, truyền cảm hứng cho sinh viên khác.

**Xây dựng hệ thống cảnh báo sớm Stress**

Theo dõi biến động giấc ngủ, thời gian sử dụng màn hình và hoạt động thể chất.

Phát hiện sớm các tín hiệu nguy cơ stress tăng cao để can thiệp kịp thời.